JOSÉ LUIS ALONSO BERROCAL

CIBERMETRÍA: Análisis de los dominios Web españoles.

UNIVERSIDAD DE SALAMANCA

RESUMEN

La cibermetría, considerándola de forma genérica como la disciplina que mide de forma cuantitativa distintos aspectos de Internet, se está consolidando, debido a la creciente necesidad de conocer el tamaño y el crecimiento del Web.

Pero además un adecuado conocimiento del Web nos va a permitir también crear mecanismos adecuados para generar técnicas que favorezcan la recuperación de información en un medio tan complejo como el Web.

Algunos trabajos preliminares, en España, realizados por documentalistas han empleado algunos aspectos cibermétricos, pero dichos trabajos han sido muy básicos, faltando un mayor desarrollo y sobre todo un estudio muestral mucho más amplio.

Se precisa de bases de datos con un tamaño muestral importante, para que los resultados que se obtengan sean significativos, para poder realizar una amplia variedad de cálculos que determinen unos resultados más fiables.

En esta tesis doctoral nos acercaremos precisamente al estudio de la cibermetría, indicando tres vías de estudio, que consideramos nos permiten analizar adecuadamente el Web.

ABSTRACT

The cibermetrics, considering it of generic form like the discipline that measures of quantitative form different aspects from Internet, is consolidating, due to the increasing necessity to know the size and the growth of the Web.

But in addition a suitable knowledge to the Web is going to also allow to us to create suitable mechanisms to generate techniques that favor the information retrieval in as complex means as the Web.

Some preliminary works, in Spain, made by documentalistas have used some cibermetrics aspects, but these works a greater development has been very basic, needing and mainly a study much more ample sample.

Precise of data bases with a so large sample important, so that the results that are obtained are significant, to be able to make an ample variety of calculations that determine more trustworthy results.

In this doctoral thesis we will indeed approach the study of the cibermetrics, indicating three channels of study, that we considered allow us to analyze the Web suitably.

Índice de Contenido.

Índice de figuras.

Listado de los dominios analizados:

1. cervantes: Instituto Cervantes.
2. cicyt: Centro de Investigación en Ciencia y Tecnología.
3. ciemat: Centro de Investigaciones Energéticas, Medioambientales y tecnológicas.
4. deusto: Universidad de Deusto.
5. esa: Agencia Espacial Europea.
6. fundesco: Fundación para el Desarrollo de la Función Social de las Comunicaciones.
7. hrc: Hospital Ramón y Cajal.
8. impi: Instituto de la Pequeña y Mediana Empresa.
9. uam: Universidad Autónoma de Madrid.
10. uc3m: Universidad Carlos III de Madrid.
11. ugr: Universidad de Granada.
12. ujaen: Universidad de jaén.
13. uji: Universidad Jaume I de Castellón.
14. um: Universidad de Murcia.
15. unex: Universidad de Extremadura.
16. unican: Univesidad de Cantabria.
17. unnet: Universidad Antonio de Nebrija.
18. upco: Universidad Pontificia de Comillas.
19. upna: Universidad Pública de Navarra.
20. upsa: Universidad Pontificia de Salamanca.
21. url: Universidad Ramón Llul.
22. urv: Universidad Rovira i Virgili.
23. us: Universidad de Sevilla.
24. usal: Universidad de Salamanca.
25. uv: Universidad de Valencia.
26. uva: Universidad de Valladolid.
27. vhebron: Hospital Vall d'Hebron.

Agradecimientos.

En la elaboración de una Tesis Doctoral se cuenta con la ayuda de muchas personas que de un modo u otro favorecen la realización de la investigación y contribuyen directa o indirectamente en su conclusión.

En primer lugar quiero agradecer a mi director de tesis, el Doctor Carlos García-Figuerola Paniagua todo su apoyo y ayuda en la realización de esta tesis, así como su apoyo durante todos estos años en los que hemos trabajado juntos, por tu amistad y enseñanzas. Gracias por todo.

Quiero también agradecer a Ángel Francisco Zazo Rodríguez su ayuda. Descargándome de trabajo, me ha permitido en los últimos tiempos dedicarme a la tesis y no pensar en algunos de los proyectos en los que estamos trabajando.

Agradecer a todos los compañeros de departamento que han encauzado en algún momento el rumbo perdido y que me han facilitado soporte técnico en momentos muy oportunos.

Agradecer la amabilidad que algunos autores han tenido conmigo, respondiendo a mis preguntas y permitiéndome clarificar mis ideas. Quiero destacar a los profesores Peter Willet, del *Department of Information Studies* de la Universidad de Sheffield; Wallace C. Koehler, de la *School of Library and Information Studies* de la Universidad de Oklahoma; Alan F. Smeaton de la *School of Computer Applications* de la Universidad de la Ciudad de Dublín.

Introducción y objetivos.

Introducción.

La cibermetría, considerándola de forma genérica como la disciplina que mide de forma cuantitativa distintos aspectos de Internet (Aguillo, 2000b), se está consolidando, debido a la creciente necesidad de conocer el tamaño y el crecimiento del Web.

Pero además un adecuado conocimiento del Web nos va a permitir también crear mecanismos adecuados para generar técnicas que favorezcan la recuperación de información en el Web.

Algunos trabajos preliminares, en España, realizados por documentalistas (Termens, 1997), (López, 1998), (Arellano, 1999), (Castillo, 1999) y (Codina, 2000) han empleado algunos aspectos cibermétricos, pero dichos trabajos han sido muy básicos, faltando un mayor desarrollo y sobre todo un estudio muestral mucho más amplio.

Se precisa de bases de datos con un tamaño muestral importante, para que los resultados que se obtengan sean significativos, para poder realizar una amplia variedad de cálculos que determinen unos resultados más fiables.

En esta tesis doctoral nos acercaremos precisamente al estudio de la cibermetría, indicando tres vías de estudio, que consideramos nos permiten analizar adecuadamente el Web.

El trabajo se divide en los siguientes capítulos:

Capítulo 1. *Estudio del Web*. En este capítulo se realiza una breve historia de Internet, finalizando con unos datos estadísticos de su crecimiento; una historia del Web, con los datos estadísticos de su evolución, significando la importancia que está adquiriendo en nuestra sociedad; posteriormente veremos una historia breve de la bibliometría para pasar a continuación a definir y centrarnos en la cibermetría; finalmente nos centraremos en el proceso de la investigación realizada, realizando las indicaciones necesarias para comprender como se ha realizado la investigación y los objetivos que se pretenden cubrir con esta investigación.

Capítulo 2. *Análisis cuantitativo*. En este capítulo veremos los diferentes indicadores que podemos tener a nuestro alcance para analizar el Web.

Capítulo 3. *Medidas topológicas*. Veremos una perspectiva interesante para el estudio y análisis del Web basándose en el estudio de su topología. Ha sido poco aplicada, posiblemente por los grandes requisitos de computación que se precisan y por necesitar una recogida de datos que permita realizar este tratamiento, y hasta la fecha por los sistemas empleados para la cibermetría no era posible.

Capítulo 4. *Leyes de exponenciación*. Nos centramos en una nueva línea de investigación, muy reciente, y con escaso estudio, pero que abre nuevas perspectivas al estudio del Web. Aplicaremos dichas leyes al caso del Web español, obteniendo unos nuevos datos de comparación, respecto a los dos trabajos más importantes publicados sobre esta nueva línea.

Capítulo 5. *Conclusiones.*

Capítulo 6. *Bibliografía.*

Objetivos.

Los objetivos que se pretenden con esta investigación son los siguientes:

- Plantear un procedimiento de investigación cibermétrica, empleando para ello un sistema de recogida de datos elaborado por nosotros, que nos permita realizar cualquier tipo de cálculo que precisemos y eliminando por ello los inconvenientes que plantean los sistemas que se están utilizando hasta este momento.

- Plantear tres posibles líneas de trabajo, que en conjunto ofrezcan una nueva visión de los estudios cibermétricos y que hasta este momento no se han planteado.

- Plantear un análisis cuantitativo en nuestro estudio que nos permita obtener una serie de indicadores que nos den una medida de la evolución de los dominios Web españoles. Realizaremos un estudio de los análisis planteados hasta el momento y las soluciones que nosotros hemos ofrecido para algunos de los indicadores planteados.

- Estudiaremos el Web como si fuese un grafo, según indican algunas de las teorías y realizaremos desde esa perspectiva algunos cálculos que creemos nos van a permitir analizar la topología de los dominios y ofreciendo unos valores adecuados que nos permitirán analizar la evolución de dichos dominios. Se realizará la comparación de algunos cálculos que en teoría miden lo mismo e intentaremos observar si esto es realmente así o por el contrario tienen algún comportamiento diferenciador.

- Abordaremos una nueva tendencia de estudio muy reciente, denominada leyes de exponenciación, e intentaremos aplicar dichas leyes a los dominios objeto del estudio analizando las características que estos dominios puedan plantear en función de los cálculos realizados.

- Debemos valorar si la metodología utilizada y las líneas de investigación que se abordan han sido adecuadas y obtenemos resultados válidos o resultados que puedan ser prometedores, pero que precisan de una mayor investigación y desarrollo.

1. Estudio del Web.

En el estudio del World Wide Web, creemos importante analizar, aunque sea de forma breve, algunos de los elementos que forman parte de dicho estudio. Por ello vamos a incorporar en este apartado una breve historia de Internet, que nos ofrezca los antecedentes del mismo, así como ver la evolución importantísima que está teniendo.

La historia de Internet aunque amplia y plagada de acontecimientos vamos a abordarla desde una perspectiva de síntesis, centrándonos en algunos de los aspectos que consideramos más relevantes. Esta historia que incluimos está basada en los trabajos de (Leiner, 1997), y en los publicados en la revista Novática (Leiner, 1997b), (Leiner, 1998) que pueden consultarse en la dirección Web http://www.ati.es/DOCS/internet/histint/

Se basa también en los trabajos de (Cerf, 1993), (Hardy, 1993), (Hardy, 1996), (Hauben, 1995), (Kulikowski, 1999), (Quarterman, 1990), (ARPANET, 1991) y (Hobbes, 2000)

La importancia de esta historia radica en que sus autores son los principales protagonistas de esa misma historia ofreciendo una visión de primera mano de los acontecimientos.

También debemos dar una breve historia del Web, para ver algunas de sus características más importantes, mostrando la evolución en el empleo del Web, que demuestran su cada vez mayor uso y que lo están convirtiendo en un servicio esencial para la difusión de información.

Veremos un breve recorrido por la historia de la bibliometría, centrándonos en algunos de los aspectos más relevantes, viendo diferentes definiciones, así como su alcance, para enlazar con la cibermetría y ver su definición y evolución.

Finalmente explicaremos las condiciones de nuestra investigación y comentaremos algunas de las particularidades de la misma.

1.1. Internet. Breve historia.

Internet ha supuesto una revolución sin precedentes en el mundo de la informática y de las comunicaciones. Los inventos del telégrafo, teléfono, radio y

ordenador sentaron las bases para esta integración de capacidades nunca antes vivida. Internet es a la vez una oportunidad de difusión mundial, un mecanismo de propagación de la información y un medio de colaboración e interacción entre los individuos y sus ordenadores independientemente de su localización geográfica.

Internet representa uno de los ejemplos más exitosos de los beneficios de la inversión sostenida y del compromiso de investigación y desarrollo en infraestructuras informáticas. A raíz de la primitiva investigación en conmutación de paquetes, el gobierno, la industria y el mundo académico han sido copartícipes de la evolución y desarrollo de esta nueva y excitante tecnología.

Esta historia gira en torno a cuatro aspectos distintos. Existe una evolución tecnológica que comienza con la primitiva investigación en conmutación de paquetes, ARPANET (*Advanced Research Projects Agency Network*) y tecnologías relacionadas en virtud de la cual la investigación actual continúa tratando de expandir los horizontes de la infraestructura en dimensiones tales como escala, rendimiento y funcionalidades de alto nivel. Hay aspectos de operación y gestión de una infraestructura operacional global y compleja. Existen aspectos sociales, que tuvieron como consecuencia el nacimiento de una amplia comunidad de internautas trabajando juntos para crear y hacer evolucionar la tecnología. Y finalmente, el aspecto de comercialización que desemboca en una transición enormemente efectiva desde los resultados de la investigación hacia una infraestructura informática ampliamente desarrollada y disponible.

Internet hoy en día es una infraestructura informática ampliamente extendida. Su primer prototipo es a menudo denominado *National Global or Galactic Information Infrastructure* (Infraestructura de Información Nacional Global o Galáctica). Su historia es compleja y comprende muchos aspectos: tecnológicos, de organización y comunitarios. Y su influencia alcanza no solamente al campo técnico de las comunicaciones computacionales sino también a toda la sociedad en la medida en que nos movemos hacia el incremento del uso de las herramientas *online* para llevar a cabo el comercio electrónico, la adquisición de información y la acción en comunidad.

1.1.1. Orígenes de Internet

La primera descripción documentada acerca de las interacciones sociales que podrían ser propiciadas a través del *networking* (trabajo en red) está contenida en una serie de memorándums escritos por (Licklider, 1962), del Massachusetts Institute of Technology (MIT), en Agosto de 1962, en los cuales Licklider discute sobre su concepto de *Galactic Network* (Red Galáctica). Él concibió una red interconectada globalmente a través de la cual cada uno pudiera acceder desde cualquier lugar a datos y programas. En esencia, el concepto era muy parecido a la Internet actual. Licklider fue el principal responsable del programa de investigación en ordenadores de la DARPA (*Defence Advanced Research Project Agency*) desde Octubre de 1962. Mientras trabajó en DARPA convenció a sus sucesores Ivan Sutherland, Bob Taylor, y el investigador del MIT Lawrence G. Roberts de la importancia del concepto de trabajo en red.

En Julio de 1961 Leonard Kleinrock (Kleinrock, 1961) publicó desde el MIT el primer documento sobre la teoría de conmutación de paquetes. Kleinrock convenció a Roberts de la factibilidad teórica de las comunicaciones vía paquetes en lugar de circuitos, lo cual resultó ser un gran avance en el camino hacia el trabajo informático en red. El otro paso fundamental fue hacer dialogar a los ordenadores entre sí. Para explorar este terreno, en 1965, Roberts conectó un ordenador TX2 en Massachusetts con un Q-32 en California a través de una línea telefónica conmutada de baja velocidad, creando así la primera (aunque reducida) red de ordenadores de área amplia jamás construida. El resultado del experimento fue la constatación de que los ordenadores de tiempo compartido podían trabajar juntos correctamente, ejecutando programas y recuperando datos a discreción en la máquina remota, pero que el sistema telefónico de conmutación de circuitos era totalmente inadecuado para esta labor. La convicción de Kleinrock acerca de la necesidad de la conmutación de paquetes quedó pues confirmada.

A finales de 1966 Roberts se trasladó a la DARPA a desarrollar el concepto de red de ordenadores y rápidamente confeccionó su plan para ARPANET, publicándolo en 1967. En la conferencia en la que presentó el documento (Roberts, 1966) se exponía también un trabajo sobre el concepto de red de paquetes a cargo de Donald Davies y Roger Scantlebury del *National Physical Laboratory* (NPL).

Scantlebury le habló a Roberts sobre su trabajo en el NPL así como sobre el de Paul Baran (Baran, 1964) y otros en RAND. El grupo RAND había escrito un documento sobre redes de conmutación de paquetes para comunicación vocal segura en el ámbito militar, en 1964. Ocurrió que los trabajos del MIT (1961-67), RAND (1962-65) y NPL (1964-67) habían discurrido en paralelo sin que los investigadores hubieran conocido el trabajo de los demás. La palabra *packet* (paquete) fue adoptada a partir del trabajo del NPL y la velocidad de la línea propuesta para ser usada en el diseño de ARPANET fue aumentada desde 2,4 Kbps hasta 50 Kbps.

En Agosto de 1968, después de que Roberts y la comunidad de la DARPA hubieran refinado la estructura global y las especificaciones de ARPANET, DARPA lanzó un RFQ para el desarrollo de uno de sus componentes clave: los conmutadores de paquetes llamados *interface message processors* (IMPs, procesadores de mensajes de interfaz). El RFQ fue ganado en Diciembre de 1968 por un grupo encabezado por Frank Heart, de Bolt Beranek y Newman (BBN). Así como el equipo de BBN trabajó en IMPs con Bob Kahn tomando un papel principal en el diseño de la arquitectura de la ARPANET global, la topología de red y el aspecto económico fueron diseñados y optimizados por Roberts trabajando con Howard Frank y su equipo en la Network Analysis Corporation, y el sistema de medida de la red fue preparado por el equipo de Kleinrock de la Universidad de California, en Los Angeles.

A causa del temprano desarrollo de la teoría de conmutación de paquetes de Kleinrock y su énfasis en el análisis, diseño y medición, su *Network Measurement Center* (Centro de Medidas de Red) en la UCLA (Universidad de California Los Ángeles) fue seleccionado para ser el primer nodo de ARPANET. Todo ello ocurrió en Septiembre de 1969, cuando BBN instaló el primer IMP en la UCLA y quedó conectado el primer ordenador *host*. El proyecto de Doug Engelbart denominado *Augmentation of Human Intelect* (Aumento del Intelecto Humano) que incluía NLS, un primitivo sistema hipertexto en el Instituto de Investigación de Standford (SRI) proporcionó un segundo nodo. El SRI patrocinó el *Network Information Center*, liderado por Elizabeth (Jake) Feinler, que desarrolló funciones tales como mantener tablas de nombres de *host* para la traducción de direcciones así como un directorio de RFCs (*Request For Comments*). Un mes más tarde, cuando el SRI fue conectado a ARPANET, el primer mensaje de *host* a *host* fue

enviado desde el laboratorio de Leinrock al SRI. Se añadieron dos nodos en la Universidad de California, Santa Bárbara, y en la Universidad de Utah. Estos dos últimos nodos incorporaron proyectos de visualización de aplicaciones, con Glen Culler y Burton Fried en la UCSB (Universidad de California Santa Bárbara) investigando métodos para mostrar funciones matemáticas mediante el uso de *"storage displays"* (mecanismos que incorporan *buffers* de monitorización distribuidos en red para facilitar el refresco de la visualización) para tratar con el problema de refrescar sobre la red, y Robert Taylor y Ivan Sutherland en Utah investigando métodos de representación en 3-D a través de la red. Así, a finales de 1969, cuatro ordenadores *host* fueron conectados conjuntamente a la ARPANET inicial y se hizo realidad una embrionaria Internet. Incluso en esta primitiva etapa, hay que reseñar que la investigación incorporó tanto el trabajo mediante la red ya existente como la mejora de la utilización de dicha red. Esta tradición continúa hasta el día de hoy.

Se siguieron conectando ordenadores rápidamente a la ARPANET durante los años siguientes y el trabajo continuó para completar un protocolo *host* a *host* funcionalmente completo, así como software adicional de red. En Diciembre de 1970, el *Network Working Group* (NWG) liderado por S.Crocker acabó el protocolo *host* a *host* inicial para ARPANET, llamado *Network Control Protocol* (NCP, protocolo de control de red). Cuando en los nodos de ARPANET se completó la implementación del NCP durante el periodo 1971-72, los usuarios de la red pudieron finalmente comenzar a desarrollar aplicaciones.

En Octubre de 1972, Kahn organizó una gran y muy exitosa demostración de ARPANET en la *International Computer Communication Conference* (Kahn, 1972). Esta fue la primera demostración pública de la nueva tecnología de red. Fue también en 1972 cuando se introdujo la primera aplicación "estrella": el correo electrónico.

En Marzo, Ray Tomlinson, de BBN, escribió el software básico de envío-recepción de mensajes de correo electrónico, impulsado por la necesidad que tenían los desarrolladores de ARPANET de un mecanismo sencillo de coordinación. En Julio, Roberts expandió su valor añadido escribiendo el primer programa de utilidad de correo electrónico para relacionar, leer selectivamente, almacenar,

reenviar y responder a mensajes. Desde entonces, la aplicación de correo electrónico se convirtió en la mayor de la red durante más de una década. Fue precursora del tipo de actividad que observamos hoy día en la *World Wide Web*, es decir, del enorme crecimiento de todas las formas de tráfico persona a persona.

1.1.2. Conceptos iniciales sobre *Internetting*

La ARPANET original evolucionó hacia Internet. Internet se basó en la idea de que habría múltiples redes independientes, de diseño casi arbitrario, empezando por ARPANET como la red pionera de conmutación de paquetes, pero que pronto incluiría redes de paquetes por satélite, redes de paquetes por radio y otros tipos de red. Internet como ahora la conocemos encierra una idea técnica clave, la de arquitectura abierta de trabajo en red. Bajo este enfoque, la elección de cualquier tecnología de red individual no respondería a una arquitectura específica de red sino que podría ser seleccionada libremente por un proveedor e interactuar con las otras redes a través del metanivel de la arquitectura de *Internetworking* (trabajo entre redes). Hasta ese momento, había un sólo método para "federar" redes. Era el tradicional método de conmutación de circuitos, por el cual las redes se interconectaban a nivel de circuito pasándose bits individuales síncronamente a lo largo de una porción de circuito que unía un par de sedes finales. Cabe recordar que Kleinrock había mostrado en 1961 que la conmutación de paquetes era el método de conmutación más eficiente. Juntamente con la conmutación de paquetes, las interconexiones de propósito especial entre redes constituían otra posibilidad. Y aunque había otros métodos limitados de interconexión de redes distintas, éstos requerían que una de ellas fuera usada como componente de la otra en lugar de actuar simplemente como un extremo de la comunicación para ofrecer servicio *end-to-end* (extremo a extremo).

En una red de arquitectura abierta, las redes individuales pueden ser diseñadas y desarrolladas separadamente y cada una puede tener su propia y única interfaz, que puede ofrecer a los usuarios y/u otros proveedores, incluyendo otros proveedores de Internet. Cada red puede ser diseñada de acuerdo con su entorno específico y los requerimientos de los usuarios de aquella red. No existen generalmente restricciones en los tipos de red que pueden ser incorporadas ni tampoco en su ámbito geográfico, aunque ciertas consideraciones pragmáticas

determinan qué posibilidades tienen sentido. La idea de arquitectura de red abierta fue introducida primeramente por Kahn un poco antes de su llegada a la DARPA en 1972. Este trabajo fue originalmente parte de su programa de paquetería por radio, pero más tarde se convirtió por derecho propio en un programa separado. Entonces, el programa fue llamado *Internetting*. La clave para realizar el trabajo del sistema de paquetería por radio fue un protocolo extremo a extremo seguro que pudiera mantener la comunicación efectiva frente a los cortes e interferencias de radio y que pudiera manejar las pérdidas intermitentes como las causadas por el paso a través de un túnel o el bloqueo a nivel local. Kahn pensó primero en desarrollar un protocolo local sólo para la red de paquetería por radio porque ello le hubiera evitado tratar con la multitud de sistemas operativos distintos y continuar usando NCP.

Sin embargo, NCP no tenía capacidad para direccionar redes y máquinas más allá de un destino IMP en ARPANET y de esta manera se requerían ciertos cambios en el NCP. La premisa era que ARPANET no podía ser cambiado en este aspecto. El NCP se basaba en ARPANET para proporcionar seguridad extremo a extremo. Si alguno de los paquetes se perdía, el protocolo y presumiblemente cualquier aplicación soportada sufriría una grave interrupción. En este modelo, el NCP no tenía control de errores en el *host* porque ARPANET había de ser la única red existente y era tan fiable que no requería ningún control de errores en la parte de los *host*s.

Así, Kahn decidió desarrollar una nueva versión del protocolo que pudiera satisfacer las necesidades de un entorno de red de arquitectura abierta. El protocolo podría eventualmente ser denominado *"transmision-control protocol/Internet protocol"* (TCP/IP, protocolo de control de transmisión /protocolo de Internet). Así como el NCP tendía a actuar como un *driver* (manejador) de dispositivo, el nuevo protocolo sería más bien un protocolo de comunicaciones.

1.1.3. Reglas clave

Cuatro fueron las reglas fundamentales en las primeras ideas de Kahn:

1) Cada red distinta debería mantenerse por sí misma y no deberían requerirse cambios internos a ninguna de ellas para conectarse a Internet.

2) Las comunicaciones deberían ser establecidas en base a la filosofía del "*best-effort*" (lo mejor posible). Si un paquete no llegara a su destino debería ser en breve retransmitido desde el emisor.

3) Para interconectar redes se usarían cajas negras, las cuales más tarde serían denominadas *gateways* (pasarelas) y *routers* (enrutadores). Los gateways no deberían almacenar información alguna sobre los flujos individuales de paquetes que circulasen a través de ellos, manteniendo de esta manera su simplicidad y evitando la complicada adaptación y recuperación a partir de las diversas modalidades de fallo.

4) No habría ningún control global a nivel de operaciones.

Otras cuestiones clave que debían ser resueltas eran:

- Algoritmos para evitar la pérdida de paquetes en base a la invalidación de las comunicaciones y la reiniciación de las mismas para la retransmisión exitosa desde el emisor.
- Provisión de *pipelining* ("tuberías") *host* a *host* de tal forma que se pudieran enrutar múltiples paquetes desde el origen al destino a discreción de los *host*s participantes, siempre que las redes intermedias lo permitieran.
- Funciones de pasarela para permitir redirigir los paquetes adecuadamente. Esto incluía la interpretación de las cabeceras IP para enrutado, manejo de interfaces y división de paquetes en trozos más pequeños si fuera necesario.
- La necesidad de controles (*checksums*) extremo a extremo, reensamblaje de paquetes a partir de fragmentos, y detección de duplicados si los hubiere.
- Necesidad de direccionamiento global.
- Técnicas para el control del flujo *host* a *host*.
- Interacción con varios sistemas operativos.

- Implementación eficiente y rendimiento de la red, aunque en principio éstas eran consideraciones secundarias.

Kahn empezó a trabajar en un conjunto de principios para sistemas operativos orientados a comunicaciones mientras se encontraba en BBN y escribió algunas de sus primeras ideas en un memorándum interno de BBN titulado "*Communications Principles for Operating Systems*". En ese momento, se dió cuenta de que le sería necesario aprender los detalles de implementación de cada sistema operativo para tener la posibilidad de incluir nuevos protocolos de manera eficiente. Así, en la primavera de 1973, después de haber empezado el trabajo de "Internetting", le pidió a Vinton Cerf (entonces en la Universidad de Stanford) que trabajara con él en el diseño detallado del protocolo. Cerf había estado íntimamente implicado en el diseño y desarrollo original del NCP y ya tenía conocimientos sobre la construcción de interfaces con los sistemas operativos existentes. De esta forma, valiéndose del enfoque arquitectural de Kahn en cuanto a comunicaciones y de la experiencia en NCP de Cerf, se asociaron para abordar los detalles de lo que acabaría siendo TCP/IP.

El trabajo en común fue altamente productivo y la primera versión escrita (Cerf, 1974) bajo este enfoque fue distribuida en una sesión especial del INWG (*International Network Working Group*, Grupo de trabajo sobre redes internacionales) que había sido convocada con motivo de una conferencia de la Universidad de Sussex en Septiembre de 1973. Cerf había sido invitado a presidir el grupo y aprovechó la ocasión para celebrar una reunión de los miembros del INWG, ampliamente representados en esta conferencia de Sussex.

Estas son las directrices básicas que surgieron de la colaboración entre Kahn y Cerf:

- Las comunicaciones entre dos procesos consistirían lógicamente en un larga corriente de bytes; ellos los llamaban "octetos". La posición de un octeto dentro de esta corriente de datos sería usada para identificarlo.

- El control del flujo se realizaría usando ventanas deslizantes y *acks* (abreviatura de *acknowledgment*, acuse de recibo). El destinatario

podría decidir cuando enviar acuse de recibo y cada *ack* devuelto correspondería a todos los paquetes recibidos hasta el momento.

- Se dejó abierto el modo exacto en que emisor y destinatario acordarían los parámetros sobre los tamaños de las ventanas a usar. Se usaron inicialmente valores por defecto.

- Aunque en aquellos momentos Ethernet estaba en desarrollo en el PARC de Xerox, la proliferación de LANs no había sido prevista entonces y mucho menos la de PCs y estaciones de trabajo. El modelo original fue concebido como un conjunto, que se esperaba reducido, de redes de ámbito nacional tipo ARPANET. De este modo, se usó una dirección IP de 32 bits, de la cual los primeros 8 identificaban la red y los restantes 24 designaban el *host* dentro de dicha red. La decisión de que 256 redes sería suficiente para el futuro previsible debió empezar a reconsiderarse en cuanto las LANs empezaron a aparecer a finales de los setenta.

El documento original de Cerf y Kahn sobre Internet describía un protocolo, llamado TCP, que se encargaba de proveer todos los servicios de transporte y reenvío en Internet. Kahn pretendía que TCP diera soporte a un amplio rango de servicios de transporte, desde el envío secuencial de datos, totalmente fiable (modelo de circuito virtual) hasta un servicio de datagramas en el que la aplicación hiciera un uso directo del servicio de red subyacente, lo que podría implicar pérdida ocasional, corrupción o reordenación de paquetes.

Sin embargo, el esfuerzo inicial de implementación de TCP dio lugar a una versión que sólo permitía circuitos virtuales. Este modelo funcionaba perfectamente en la transferencia de ficheros y en las aplicaciones de *login* remoto, pero algunos de los primeros trabajos sobre aplicaciones avanzadas de redes (en particular el empaquetamiento de voz en los años 70) dejó bien claro que, en ciertos casos, el TCP no debía encargarse de corregir las pérdidas de paquetes y que había que dejar a la aplicación que se ocupara de ello. Esto llevó a la reorganización del TCP original en dos protocolos: uno sencillo, IP, que se encargara tan sólo de dar una dirección a los paquetes y de reenviarlos; y un TCP que se dedicara a una serie de funcionalidades como el control del flujo y la recuperación de los paquetes

perdidos. Para aquellas aplicaciones que no precisan los servicios de TCP, se añadió un protocolo alternativo llamado UDP (*User Datagram Protocol*, protocolo de datagramas de usuario) dedicado a dar un acceso directo a los servicios básicos del IP.

Una de las motivaciones iniciales de ARPANET e Internet fue compartir recursos, por ejemplo, permitiendo que usuarios de redes de paquetes sobre radio pudieran acceder a sistemas de tiempo compartido conectados a ARPANET. Conectar las dos redes era mucho más económico que duplicar estos carísimos ordenadores. Sin embargo, mientras la transferencia de ficheros y el *login* remoto (Telnet) eran aplicaciones muy importantes, de todas las de esta época probablemente sea el correo electrónico la que haya tenido un impacto más significativo. El correo electrónico dio lugar a un nuevo modelo de comunicación entre las personas y cambió la naturaleza de la colaboración. Su influencia se manifestó en primer lugar en la construcción de la propia Internet (como veremos más adelante), y posteriormente, en buena parte de la sociedad.

Se propusieron otras aplicaciones en los primeros tiempos de Internet, desde la comunicación vocal basada en paquetes (precursora de la telefonía sobre Internet) o varios modelos para compartir ficheros y discos, hasta los primeros "programas-gusano" que mostraban el concepto de agente (y, por supuesto, de virus). Un concepto clave en Internet es que no fue diseñada para una única aplicación sino como una infraestructura general dentro de la que podrían concebirse nuevos servicios, como con posterioridad demostró la aparición de la *World Wide Web*. Este fue posible solamente debido a la orientación de propósito general que tenía el servicio implementado mediante TCP e IP.

1.1.4. Ideas a prueba

DARPA formalizó tres contratos con Stanford (Cerf), BBN (Ray Tomlinson) y UCLA (Peter Kirstein) para implementar TCP/IP (en el documento original de Cerf y Kahn se llamaba simplemente TCP pero contenía ambos componentes). El equipo de Stanford, dirigido por Cerf, produjo las especificaciones detalladas y al cabo de un año hubo tres implementaciones independientes de TCP que podían interoperar.

Este fue el principio de un largo periodo de experimentación y desarrollo para evolucionar y madurar el concepto y tecnología de Internet. Partiendo de las tres primeras redes ARPANET, radio y satélite y de sus comunidades de investigación iniciales, el entorno experimental creció hasta incorporar esencialmente cualquier forma de red y una amplia comunidad de investigación y desarrollo. Cada expansión afrontó nuevos desafíos.

Las primeras implementaciones de TCP se hicieron para grandes sistemas en tiempo compartido como Tenex y TOPS 20. Cuando aparecieron los ordenadores de sobremesa (*desktop*), TCP era demasiado grande y complejo como para funcionar en ordenadores personales. David Clark y su equipo de investigación del MIT empezaron a buscar la implementación de TCP más sencilla y compacta posible. La desarrollaron, primero para el Alto de Xerox (la primera estación de trabajo personal desarrollada en el PARC de Xerox), y luego para el PC de IBM. Esta implementación operaba con otras de TCP, pero estaba adaptada al conjunto de aplicaciones y a las prestaciones de un ordenador personal, y demostraba que las estaciones de trabajo, al igual que los grandes sistemas, podían ser parte de Internet.

En los años 80, el desarrollo de LAN, PC y estaciones de trabajo permitió que la naciente Internet floreciera. La tecnología Ethernet, desarrollada por Bob Metcalfe en el PARC de Xerox en 1973, es la dominante en Internet, y los PCs y las estaciones de trabajo los modelos de ordenador dominantes. El cambio que supone pasar de una pocas redes con un modesto número de *hosts* (el modelo original de ARPANET) a tener muchas redes dio lugar a nuevos conceptos y a cambios en la tecnología. En primer lugar, hubo que definir tres clases de redes (A, B y C) para acomodar todas las existentes. La clase A representa a las redes grandes, a escala nacional (pocas redes con muchos ordenadores); la clase B representa redes regionales; por último, la clase C representa redes de área local (muchas redes con relativamente pocos ordenadores).

Como resultado del crecimiento de Internet, se produjo un cambio de gran importancia para la red y su gestión. Para facilitar el uso de Internet por sus usuarios se asignaron nombres a los *host*s de forma que resultara innecesario recordar sus direcciones numéricas. Originalmente había un número muy limitado de máquinas,

por lo que bastaba con una simple tabla con todos los ordenadores y sus direcciones asociadas.

El cambio hacia un gran número de redes gestionadas independientemente (por ejemplo, las LAN) significó que no resultara ya fiable tener una pequeña tabla con todos los *host*s. Esto llevó a la invención del DNS (*Domain Name System*, sistema de nombres de dominio) por Paul Mockapetris de USC/ISI. El DNS permitía un mecanismo escalable y distribuido para resolver jerárquicamente los nombres de los *host*s (por ejemplo, *www.acm.org* o *www.ati.es*) en direcciones de Internet.

El incremento del tamaño de Internet resultó también un desafío para los *routers*. Originalmente había un sencillo algoritmo de enrutamiento que estaba implementado uniformemente en todos los routers de Internet. A medida que el número de redes en Internet se multiplicaba, el diseño inicial no era ya capaz de expandirse, por lo que fue sustituido por un modelo jerárquico de enrutamiento con un protocolo IGP (*Interior Gateway Protocol*, protocolo interno de pasarela) usado dentro de cada región de Internet y un protocolo EGP (*Exterior Gateway Protocol*, protocolo externo de pasarela) usado para mantener unidas las regiones. El diseño permitía que distintas regiones utilizaran IGP distintos, por lo que los requisitos de coste, velocidad de configuración, robustez y escalabilidad, podían ajustarse a cada situación. Los algoritmos de enrutamiento no eran los únicos en poner en dificultades la capacidad de los *routers*, también lo hacía el tamaño de la tablas de direccionamiento. Se presentaron nuevas aproximaciones a la agregación de direcciones (en particular CIDR, *Classless Interdomain Routing*, enrutamiento entre dominios sin clase) para controlar el tamaño de las tablas de enrutamiento.

A medida que evolucionaba Internet, la propagación de los cambios en el software, especialmente el de los *host*s, se fue convirtiendo en uno de sus mayores desafíos. DARPA financió a la Universidad de California en Berkeley en una investigación sobre modificaciones en el sistema operativo Unix, incorporando el TCP/IP desarrollado en BBN. Aunque posteriormente Berkeley modificó esta implementación del BBN para que operara de forma más eficiente con el sistema y el kernel de Unix, la incorporación de TCP/IP en el sistema Unix BSD demostró ser un elemento crítico en la difusión de los protocolos entre la comunidad investigadora. BSD empezó a ser utilizado en sus operaciones diarias por buena

parte de la comunidad investigadora en temas relacionados con informática. Visto en perspectiva, la estrategia de incorporar los protocolos de Internet en un sistema operativo utilizado por la comunidad investigadora fue uno de los elementos clave en la exitosa y amplia aceptación de Internet.

Uno de los desafíos más interesantes fue la transición del protocolo para *host*s de ARPANET desde NCP a TCP/IP el 1 de enero de 1983. Se trataba de una ocasión muy importante que exigía que todos los *host*s se convirtieran simultáneamente o que permanecieran comunicados mediante mecanismos desarrollados para la ocasión. La transición fue cuidadosamente planificada dentro de la comunidad con varios años de antelación a la fecha, pero fue sorprendentemente sobre ruedas (a pesar de dar la lugar a la distribución de insignias con la inscripción "Yo sobreviví a la transición a TCP/IP").

TCP/IP había sido adoptado como un estándar por el ejército norteamericano tres años antes, en 1980. Esto permitió al ejército empezar a compartir la tecnología DARPA basada en Internet y llevó a la separación final entre las comunidades militares y no militares. En 1983 ARPANET estaba siendo usada por un número significativo de organizaciones operativas y de investigación y desarrollo en el área de la defensa. La transición desde NCP a TCP/IP en ARPANET permitió la división en una MILNET para dar soporte a requisitos operativos y una ARPANET para las necesidades de investigación.

Así, en 1985, Internet estaba firmemente establecida como una tecnología que ayudaba a una amplia comunidad de investigadores y desarrolladores, y empezaba a ser empleada por otros grupos en sus comunicaciones diarias entre ordenadores. El correo electrónico se empleaba ampliamente entre varias comunidades, a menudo entre distintos sistemas. La interconexión entre los diversos sistemas de correo demostraba la utilidad de las comunicaciones electrónicas entre personas.

1.1.5. La transición hacia una infraestructura global

Al mismo tiempo que la tecnología Internet estaba siendo validada experimentalmente y usada ampliamente entre un grupo de investigadores de informática se estaban desarrollando otras redes y tecnologías. La utilidad de las

redes de ordenadores (especialmente el correo electrónico utilizado por los contratistas de DARPA y el Departamento de Defensa en ARPANET) siguió siendo evidente para otras comunidades y disciplinas de forma que a mediados de los años 70 las redes de ordenadores comenzaron a difundirse allá donde se podía encontrar financiación para las mismas. El Departamento norteamericano de Energía (DoE, *Deparment of Energy*) estableció MFENet (*Magnetic Fusion Energy NETwork*) para sus investigadores que trabajaban sobre energía de fusión, mientras que los físicos de altas energías fueron los encargados de construir HEPNet (*High Energy Physics network*). Los físicos de la NASA continuaron con SPAN (*Space Physics and Analysis Network*) y Rick Adrion, David Farber y Larry Landweber fundaron CSNET (*Computer & Science Network*) para la comunidad informática académica y de la industria con la financiación inicial de la NFS (*National Science Foundation*, Fundación Nacional de la Ciencia) de Estados Unidos. La libre diseminación del sistema operativo Unix de ATT dio lugar a USENET (*USEr's NETwork*), basada en los protocolos de comunicación UUCP de Unix, y en 1981 Greydon Freeman e Ira Fuchs diseñaron BITNET (*Because It's Time Network*) , que unía los ordenadores centrales del mundo académico siguiendo el paradigma de correo electrónico como "postales". Con la excepción de BITNET y USENET, todas las primeras redes (como ARPANET) se construyeron para un propósito determinado. Es decir, estaban dedicadas (y restringidas) a comunidades cerradas de estudiosos; de ahí las escasas presiones por hacer estas redes compatibles y, en consecuencia, el hecho de que durante mucho tiempo no lo fueran. Además, estaban empezando a proponerse tecnologías alternativas en el sector comercial, como XNS de Xerox, DECNet, y la SNA de IBM. Sólo restaba que los programas ingleses JANET (1984) y norteamericano NSFNET (1985) anunciaran explícitamente que su propósito era servir a toda la comunidad de la enseñanza superior sin importar su disciplina. De hecho, una de las condiciones para que una universidad norteamericana recibiera financiación de la NSF para conectarse a Internet era que "la conexión estuviera disponible para *todos* los usuarios cualificados del campus".

En 1985 Dennins Jenning acudió desde Irlanda para pasar un año en NFS dirigiendo el programa NSFNET. Trabajó con el resto de la comunidad para ayudar a la NSF a tomar una decisión crítica: si TCP/IP debería ser obligatorio en el programa NSFNET. Cuando Steve Wolff llegó al programa NFSNET en 1986

reconoció la necesidad de una infraestructura de red amplia que pudiera ser de ayuda a la comunidad investigadora y a la académica en general, junto a la necesidad de desarrollar una estrategia para establecer esta infraestructura sobre bases independientes de la financiación pública directa. Se adoptaron varias políticas y estrategias para alcanzar estos fines.

La NSF optó también por mantener la infraestructura organizativa de Internet existente (DARPA) dispuesta jerárquicamente bajo el IAB (*Internet Activities Board*, Comité de Actividades de Internet). La declaración pública de esta decisión firmada por todos sus autores (por los grupos de Arquitectura e Ingeniería de la IAB, y por el NTAG de la NSF) apareció como la RFC 985 ("Requisitos para pasarelas de Internet") que formalmente aseguraba la interoperatividad entre las partes de Internet dependientes de DARPA y de NSF.

Junto a la selección de TCP/IP para el programa NSFNET, las agencias federales norteamericanas idearon y pusieron en práctica otras decisiones que llevaron a la Internet de hoy:

Las agencias federales compartían el coste de la infraestructura común, como los circuitos transoceánicos. También mantenían la gestión de puntos de interconexión para el tráfico entre agencias: los "Federal Internet Exchanges" (FIX-E y FIX-W) que se desarrollaron con este propósito sirvieron de modelo para los puntos de acceso a red y los sistemas *IX que son unas de las funcionalidades más destacadas de la arquitectura de la Internet actual.

Para coordinar estas actividades se formó el FNC (*Federal Networking Council*, Consejo Federal de Redes). El FNC cooperaba también con otras organizaciones internacionales, como RARE en Europa, a través del CCIRN (*Coordinating Committee on Intercontinental Research Networking*, Comité de Coordinación Intercontinental de Investigación sobre Redes) para coordinar el apoyo a Internet de la comunidad investigadora mundial.

Esta cooperación entre agencias en temas relacionados con Internet tiene una larga historia. En 1981, un acuerdo sin precedentes entre Farber, actuando en nombre de CSNET y NSF, y Kahn por DARPA, permitió que el tráfico de CSNET

compartiera la infraestructura de ARPANET de acuerdo según parámetros estadísticos.

En consecuencia, y de forma similar, la NFS promocionó sus redes regionales de NSFNET, inicialmente académicas, para buscar clientes comerciales, expandiendo sus servicios y explotando las economías de escala resultantes para reducir los costes de suscripción para todos.

En el *backbone* NFSNET (el segmento que cruza los EE.UU.) NSF estableció una política aceptable de uso (AUP, *Acceptable-Use Policy*) que prohibía el uso del *backbone* para fines "que no fueran de apoyo a la Investigación y la Educación". El predecible e intencionado resultado de promocionar el tráfico comercial en la red a niveles locales y regionales era estimular la aparición y/o crecimiento de grandes redes privadas y competitivas como PSI, UUNET, ANS CO+RE, y, posteriormente, otras. Este proceso de aumento de la financiación privada para el uso comercial se resolvió tras largas discusiones que empezaron en 1988 con una serie de conferencias patrocinadas por NSF en la *Kennedy School of Government* de la Universidad de Harvard, bajo el lema "La comercialización y privatización de Internet", complementadas por la lista "*com-priv*" de la propia red.

En 1988 un comité del *National Research Council* (Consejo Nacional de Investigación), presidido por Kleinrock y entre cuyos miembros estaban Clark y Kahn, elaboró un informe dirigido a la NSF y titulado *"Towards a National Research Network"*. El informe llamó la atención del entonces senador Al Gore (Vicepresidente de los EE.UU. desde 1992) le introdujo en las redes de alta velocidad que pusieron los cimientos de la futura «Autopista de la Información».

La política de privatización de la NSF culminó en Abril de 1995 con la eliminación de la financiación del backbone NSFNET. Los fondos así recuperados fueron redistribuidos competitivamente entre redes regionales para comprar conectividad de ámbito nacional a Internet a las ahora numerosas redes privadas de larga distancia.

El *backbone* había hecho la transición desde una red construida con *routers* de la comunidad investigadora (los *routers* Fuzzball de David Mills) a equipos

comerciales. En su vida de ocho años y medio, el *backbone* había crecido desde seis nodos con enlaces de 56Kb a 21 nodos con enlaces múltiples de 45Mb.Había visto crecer Internet hasta alcanzar más de 50.000 redes en los cinco continentes y en el espacio exterior, con aproximadamente 29.000 redes en los Estados Unidos.

El efecto del ecumenismo del programa NSFNET y su financiación (200 millones de dólares entre 1986 y 1995)y de la calidad de los protocolos fue tal que en 1990, cuando la propia ARPANET se disolvió, TCP/IP había sustituido o marginado a la mayor parte de los restantes protocolos de grandes redes de ordenadores e IP estaba en camino de convertirse en *el* servicio portador de la llamada Infraestructura.

1.1.6. El papel de la documentación

Un aspecto clave del rápido crecimiento de Internet ha sido el acceso libre y abierto a los documentos básicos, especialmente a las especificaciones de los protocolos.

Los comienzos de Arpanet y de Internet en la comunidad de investigación universitaria estimularon la tradición académica de la publicación abierta de ideas y resultados. Sin embargo, el ciclo normal de la publicación académica tradicional era demasiado formal y lento para el intercambio dinámico de ideas, esencial para crear redes.

(Crocker, 1969), entonces en UCLA, dio un paso clave al establecer la serie de notas RFC (*Request For Comments*, petición de comentarios). Estos memorándums pretendieron ser una vía informal y de distribución rápida para compartir ideas con otros investigadores en redes. Al principio, las RFC fueron impresas en papel y distribuidas vía correo "lento". Pero cuando el FTP (*File Transfer Protocol*, protocolo de transferencia de ficheros) empezó a usarse, las RFC se convirtieron en ficheros difundidos *online* a los que se accedía vía FTP. Hoy en día, desde luego, están disponibles en el World Wide Web en decenas de emplazamientos en todo el mundo. SRI, en su papel como Centro de Información en la Red, mantenía los directorios *online*. Jon Postel actuaba como editor de RFC y como gestor de la administración centralizada de la asignación de los números de protocolo requeridos, tareas en las que continúa hoy en día.

El efecto de las RFC era crear un bucle positivo de realimentación, con ideas o propuestas presentadas a base de que una RFC impulsara otra RFC con ideas adicionales y así sucesivamente. Una vez se hubiera obtenido un consenso se prepararía un documento de especificación. Tal especificación seria entonces usada como la base para las implementaciones por parte de los equipos de investigación.

Con el paso del tiempo, las RFC se han enfocado a estándares de protocolo –las especificaciones oficiales- aunque hay todavía RFC informativas que describen enfoques alternativos o proporcionan información de soporte en temas de protocolos e ingeniería. Las RFC son vistas ahora como los documentos de registro dentro de la comunidad de estándares y de ingeniería en Internet.

El acceso abierto a las RFC –libre si se dispone de cualquier clase de conexión a Internet- promueve el crecimiento de Internet porque permite que las especificaciones sean usadas a modo de ejemplo en las aulas universitarias o por emprendedores al desarrollar nuevos sistemas.

El *e-mail* o correo electrónico ha supuesto un factor determinante en todas las áreas de Internet, lo que es particularmente cierto en el desarrollo de las especificaciones de protocolos, estándares técnicos e ingeniería en Internet. Las primitivas RFC a menudo presentaban al resto de la comunidad un conjunto de ideas desarrolladas por investigadores de un solo lugar. Después de empezar a usarse el correo electrónico, el modelo de autoría cambió: las RFC pasaron a ser presentadas por coautores con visiones en común, independientemente de su localización.

Las listas de correo especializadas ha sido usadas ampliamente en el desarrollo de la especificación de protocolos, y continúan siendo una herramienta importante. El IETF tiene ahora más de 75 grupos de trabajo, cada uno dedicado a un aspecto distinto de la ingeniería en Internet. Cada uno de estos grupos de trabajo dispone de una lista de correo para discutir uno o más borradores bajo desarrollo. Cuando se alcanza el consenso en el documento, éste puede ser distribuido como una RFC.

Debido a que la rápida expansión actual de Internet se alimenta por el aprovechamiento de su capacidad de promover la compartición de información,

deberíamos entender que el primer papel en esta tarea consistió en compartir la información acerca de su propio diseño y operación a través de los documentos RFC. Este método único de producir nuevas capacidades en la red continuará siendo crítico para la futura evolución de Internet.

1.1.7. Formación de la Comunidad Amplia

Internet es tanto un conjunto de comunidades como un conjunto de tecnologías y su éxito se puede atribuir tanto a la satisfacción de las necesidades básicas de la comunidad como a la utilización de esta comunidad de un modo efectivo para impulsar la infraestructura. El espíritu comunitario tiene una larga historia, empezando por la temprana ARPANET. Los investigadores de ésta red trabajaban como una comunidad cerrada para llevar a cabo las demostraciones iniciales de la tecnología de conmutación de paquetes descrita en la primera parte de este artículo.

Del mismo modo, la Paquetería por Satélite, la Paquetería por Radio y varios otros programas de investigación informática de la DARPA fueron actividades cooperativas y de contrato múltiple que, aún con dificultades, usaban cualquiera de los mecanismos disponibles para coordinar sus esfuerzos, empezando por el correo electrónico y siguiendo por la compartición de ficheros, acceso remoto y finalmente las prestaciones de la World Wide Web.

Cada uno de estos programas formaban un grupo de trabajo, empezando por el *ARPANET Network Working Group* (Grupo de Trabajo de la Red ARPANET). Dado que el único papel que ARPANET representaba era actuar como soporte de la infraestructura de los diversos programas de investigación, cuando Internet empezó a evolucionar, el Grupo de Trabajo de la Red se transformó en Grupo de Trabajo de Internet.

A finales de los 70, como reconocimiento de que el crecimiento de Internet estaba siendo acompañado por un incremento en el tamaño de la comunidad investigadora interesada y, por tanto, generando una necesidad creciente de mecanismos de coordinación, Vinton Cerf, por entonces director del programa de Internet en DARPA, formó varios grupos de coordinación: el ICB (*International Cooperation Board*, Consejo de Cooperación Internacional) presidido por Peter

Kirstein, para coordinar las actividades con los países cooperantes europeos y dedicado a la investigación en Paquetería por Satélite; el *Internet Research Group* (Grupo de Investigación en Internet), que fue un grupo inclusivo para proporcionar un entorno para el intercambio general de información; y el ICCB (*Internet Configuration Control Board* , Consejo de Control de la Configuración de Internet), presidido por Clark. El ICCB fue un grupo al que se pertenecía por invitación para asistir a Cerf en la dirección de la actividad incipiente de Internet.

En 1983, cuando Barry Leiner asumió la dirección del programa de investigación en DARPA, él y Clark observaron que el continuo crecimiento de la comunidad de Internet demandaba la reestructuración de los mecanismos de coordinación. El ICCB fue disuelto y sustituido por una estructura de equipos de trabajo, cada uno de ellos enfocado a un área específica de la tecnología, tal como los *routers* (encaminadores) o los protocolos extremo a extremo. Se creó el IAB (*Internet Architecture Board*, Consejo de la Arquitectura de Internet) incluyendo a los presidentes de los equipos de trabajo. Era, desde luego, solamente una coincidencia que los presidentes de los equipos de trabajo fueran las mismas personas que constituían el antiguo ICCB, y Clark continuó actuando como presidente.

Después de algunos cambios en la composición del IAB, Phill Gross fue nombrado presidente del revitalizado IETF (*Internet Engineering Task Force*, Equipo de Trabajo de Ingeniería de Internet), que en aquel momento era meramente un equipo de trabajo del IAB. Como mencionamos con anterioridad, en 1985 se produjo un tremendo crecimiento en el aspecto más práctico de la ingeniería de Internet. Tal crecimiento desembocó en una explosión en la asistencia a las reuniones del IETF y Gross se vio obligado a crear una subestructura en el IETF en forma de grupos de trabajo.

El crecimiento de Internet fue complementado por una gran expansión de la comunidad de usuarios. DARPA dejó de ser el único protagonista en la financiación de Internet. Además de NSFNET y de varias actividades financiadas por los gobiernos de Estados Unidos y otros países, el interés de parte del mundo empresarial había empezado a crecer. También en 1985, Kahn y Leiner abandonaron DARPA, y ello supuso un descenso significativo de la actividad de

Internet allí. Como consecuencia, el IAB perdió a su principal espónsor y poco a poco fue asumiendo el liderazgo.

El crecimiento continuó y desembocó en una subestructura adicional tanto en el IAB como en el IETF. El IETF integró grupos de trabajo en áreas y designó directores de área. El IESG (*Internet Engineering Steering Group*, Grupo de Dirección de Ingeniería de Internet) se formó con estos directores de área. El IAB reconoció la importancia creciente del IETF y reestructuró el proceso de estándares para reconocer explícitamente al IESG como la principal entidad de revisión de estándares. El IAB también se reestructuró de manera que el resto de equipos de trabajo (aparte del IETF) se agruparon en el IRTF (*Internet Research Task Force*, Equipo de Trabajo de Investigación en Internet), presidido por Postel, mientras que los antiguos equipos de trabajo pasaron a llamarse "grupos de investigación".

El crecimiento en el mundo empresarial trajo como consecuencia un incremento de la preocupación por el propio proceso de estándares. Desde primeros de los años 80 hasta hoy, Internet creció y está creciendo más allá de sus raíces originales de investigación para incluir a una amplia comunidad de usuarios y una actividad comercial creciente. Se puso un mayor énfasis en hacer el proceso abierto y justo. Esto, junto a una necesidad reconocida de dar soporte a la comunidad de Internet, condujo a la formación de la *Internet Society* en 1991, bajo los auspicios de la CNRI (*Corporation for National Research Initiatives*, Corporación para las Iniciativas de Investigación Nacionales) de Kahn y el liderazgo de Cerf, junto al de la CNRI.

En 1992 todavía se realizó otra reorganización: El *Internet Activities Board* (Consejo de Actividades de Internet) fue reorganizado y sustituyó al Consejo de la Arquitectura de Internet, operando bajo los auspicios de la Internet Society. Se definió una relación más estrecha entre el nuevo IAB y el IESG, tomando el IETF y el propio IESG una responsabilidad mayor en la aprobación de estándares. Por último, se estableció una relación cooperativa y de soporte mutuo entre el IAB, el IETF y la Internet Society, tomando esta última como objetivo la provisión de servicio y otras medidas que facilitarían el trabajo del IETF.

El reciente desarrollo y amplia difusión del World Wide Web ha formado una nueva comunidad, pues muchos de los que trabajan en la WWW no se

consideran a sí mismos como investigadores y desarrolladores primarios de la red. Se constituyó un nuevo organismo de coordinación, el W3C (*World Wide Web Consortium*). Liderado inicialmente desde el *Laboratory for Computer Science* del MIT por Tim Berners-Lee –el inventor del WWW- y Al Vezza, el W3C ha tomado bajo su responsabilidad la evolución de varios protocolos y estándares asociados con el Web.

Así pues, a través de más de dos décadas de actividad en Internet, hemos asistido a la continua evolución de las estructuras organizativas designadas para dar soporte y facilitar a una comunidad en crecimiento el trabajo colaborativo en temas de Internet.

1.1.8. Comercialización de la tecnología

La comercialización de Internet llevaba acarreada no sólo el desarrollo de servicios de red privados y competitivos sino también el de productos comerciales que implementen la tecnología Internet. A principios de los años 80 docenas de fabricantes incorporaron TCP/IP a sus productos debido a la aproximación de sus clientes a esta tecnología de redes. Desafortunadamente, carecían de información fiable sobre cómo funcionaba esta tecnología y cómo pensaban utilizarla sus clientes. Muchos lo enfocaron como la incorporación de funcionalidades que se añadían a sus propios sistemas de red: SNA, DECNet, Netware, NetBios. El Departamento de Defensa norteamericano hizo obligatorio el uso de TCP/IP en buena parte de sus adquisiciones de software pero dio pocas indicaciones a los suministradores sobre cómo desarrollar productos TCP/IP realmente útiles.

En 1985, reconociendo la falta de información y formación adecuadas, Dan Lynch, en cooperación con el IAB, organizó una reunión de tres días para **todos** los fabricantes que quisieran saber cómo trabajaba TCP/IP y qué es lo que aún no era capaz de hacer. Los ponentes pertenecían fundamentalmente a la comunidad investigadora de DARPA que había desarrollado los protocolos y los utilizaba en su trabajo diario. Alrededor de 250 fabricantes acudieron a escuchar a unos 50 inventores y experimentadores. Los resultados fueron una sorpresa para ambas partes: los fabricantes descubrieron con asombro que los inventores estaban abiertos a sugerencias sobre cómo funcionaban los sistemas (y sobre qué era lo que aún no eran capaces de hacer) y los inventores recibieron con agrado información

sobre nuevos problemas que no conocían pero que habían encontrado los fabricantes en el desarrollo y operación de nuevos productos. Así, quedó establecida un diálogo que ha durado más de una década.

Después de dos años de conferencias, cursos, reuniones de diseño y congresos, se organizó un acontecimiento especial para que los fabricantes cuyos productos funcionaran correctamente bajo TCP/IP pudieran mostrarlos conjuntamente durante tres días y demostraran lo bien que podían trabajar y correr en Internet. El primer "*Interop trade show*" nació en Septiembre de 1988. Cincuenta compañías presentaron sus productos y unos 5.000 ingenieros de organizaciones potencialmente compradoras acudieron a ver si todo funcionaba como se prometía. Y lo hizo. ¿Por qué? Porque los fabricantes habían trabajado intensamente para asegurar que sus productos interoperaban correctamente entre sí -incluso con los de sus competidores. El Interop ha crecido enormemente desde entonces y hoy en día se realiza cada año en siete lugares del mundo con una audiencia de 250.000 personas que acuden para comprobar qué productos interoperan correctamente con los demás, conocer cuáles son los últimos y para hablar sobre la tecnología más reciente.

En paralelo con los esfuerzos de comercialización amparados por las actividades del Interop, los fabricantes comenzaron a acudir a las reuniones de la IETF que se convocaban tres o cuatro veces al año para discutir nuevas ideas para extender el conjunto de protocolos relacionados con TCP/IP. Comenzaron con unos cientos de asistentes procedentes en su mayor parte del mundo académico y financiados por el sector público; actualmente estas reuniones atraen a varios miles de participantes, en su mayor parte del sector privado y financiados por éste. Los miembros de este grupo han hecho evolucionar el TCP/IP cooperando entre sí. La razón de que estas reuniones sean tan útiles es que acuden a ellas todas las partes implicadas: investigadores, usuarios finales y fabricantes.

La gestión de redes nos da un ejemplo de la beneficiosa relación entre la comunidad investigadora y los fabricantes. En los comienzos de Internet, se hacía hincapié en la definición e implementación de protocolos que alcanzaran la interoperación. A medida que crecía la red aparecieron situaciones en las que procedimientos desarrollados "ad hoc" para gestionar la red no eran capaces de

crecer con ella. La configuración manual de tablas fue sustituida por algoritmos distribuidos automatizados y aparecieron nuevas herramientas para resolver problemas puntuales. En 1987 quedó claro que era necesario un protocolo que permitiera que se pudieran gestionar remota y uniformemente los elementos de una red, como los *routers*. Se propusieron varios protocolos con este propósito, entre ellos el SNMP (*Single Network Management Protocol*, protocolo simple de gestión de red) diseñado, como su propio nombre indica, buscando la simplicidad; HEMS, un diseño más complejo de la comunidad investigadora; y CMIP, desarrollado por la comunidad OSI. Una serie de reuniones llevaron a tomar la decisión de desestimar HEMS como candidato para la estandarización, dejando que tanto SNMP como CMIP siguieran adelante con la idea que el primero fuera una solución inmediata mientras que CMIP pasara a ser una aproximación a largo plazo: el mercado podría elegir el que resultara más apropiado. Hoy SNMP se usa casi universalmente para la gestión de red.

En los últimos años hemos vivido una nueva fase en la comercialización. Originalmente, los esfuerzos invertidos en esta tarea consistían fundamentalmente en fabricantes que ofrecían productos básicos para trabajar en la red y proveedores de servicio que ofrecían conectividad y servicios básicos. Internet se ha acabado convirtiendo en una "**commodity**", un servicio de disponibilidad generalizada para usuarios finales, y buena parte de la atención se ha centrado en el uso de la GII (*Global Information Infraestructure*) para el soporte de servicios comerciales. Este hecho se ha acelerado tremendamente por la rápida y amplia adopción de visualizadores y de la tecnología del World Wide Web, permitiendo a los usuarios acceder fácilmente a información distribuida a través del mundo. Están disponibles productos que facilitan el acceso a esta información y buena parte de los últimos desarrollos tecnológicos están dirigidos a obtener servicios de información cada vez más sofisticados sobre comunicaciones de datos básicas de Internet.

1.1.9. Historia del futuro

El 24 de Octubre de 1995, el FNC (*Federal Networking Council*, Consejo Federal de la Red) aceptó unánimemente una resolución definiendo el término *Internet*. La definición se elaboró de acuerdo con personas de las áreas de Internet y los derechos de propiedad intelectual. La resolución: "el FNC acuerda que lo

siguiente refleja nuestra definición del término *Internet*. *Internet* hace referencia a un sistema global de información que (1) está relacionado lógicamente por un único espacio de direcciones global basado en el protocolo de Internet (IP) o en sus extensiones, (2) es capaz de soportar comunicaciones usando el conjunto de protocolos TCP/IP o sus extensiones u otros protocolos compatibles con IP, y (3) emplea, provee, o hace accesible, privada o públicamente, servicios de alto nivel en capas de comunicaciones y otras infraestructuras relacionadas aquí descritas".

Internet ha cambiado en sus dos décadas de existencia. Fue concebida en la era del tiempo compartido y ha sobrevivido en la era de los ordenadores personales, cliente-servidor, y los *network-computer*. Se ideó antes de que existieran las LAN, pero ha acomodado tanto a esa tecnología como a ATM y la conmutación de tramas. Ha dado soporte a un buen número de funciones desde compartir ficheros, y el acceso remoto, hasta compartir recursos y colaboración, pasando por el correo electrónico y, recientemente, el World Wide Web. Pero, lo que es más importante, comenzó como una creación de un pequeño grupo de investigadores y ha crecido hasta convertirse en un éxito comercial con miles de millones de dólares anuales en inversiones.

No se puede concluir diciendo que Internet ha acabado su proceso de cambio. Aunque es una red por su propia denominación y por su dispersión geográfica, su origen está en los ordenadores, no en la industria de la telefonía o la televisión. Puede -o mejor, debe- continuar cambiando y evolucionando a la velocidad de la industria del ordenador si quiere mantenerse como un elemento relevante. Ahora está cambiando para proveer nuevos servicios como el transporte en tiempo real con vistas a soportar, por ejemplo, audio y vídeo. La disponibilidad de redes penetrantes y omnipresentes, como Internet, junto con la disponibilidad de potencia de cálculo y comunicaciones asequibles en máquinas como los ordenadores portátiles, los PDA y los teléfonos celulares, está posibilitando un nuevo paradigma de informática y comunicaciones "nómadas".

Esta evolución nos traerá una nueva aplicación: telefonía Internet y, puede que poco después, televisión por Internet. Está permitiendo formas más sofisticadas de valoración y recuperación de costes, un requisito fundamental en la aplicación comercial. Está cambiando para acomodar una nueva generación de tecnologías

de red con distintas características y requisitos: desde ancho de banda doméstico a satélites. Y nuevos modos de acceso y nuevas formas de servicio que darán lugar a nuevas aplicaciones, que, a su vez, harán evolucionar a la propia red.

La cuestión más importante sobre el futuro de Internet no es cómo cambiará la tecnología, sino cómo se gestionará esa evolución. En este capítulo se ha contado cómo un grupo de diseñadores dirigió la arquitectura de Internet y cómo la naturaleza de ese grupo varió a medida que creció el número de partes interesadas. Con el éxito de Internet ha llegado una proliferación de inversores que tienen intereses tanto económicos como intelectuales en la red. Se puede ver en los debates sobre el control del espacio de nombres y en la nueva generación de direcciones IP una pugna por encontrar la nueva estructura social que guiará a Internet en el futuro. Será difícil encontrar la forma de esta estructura dado el gran número de intereses que concurren en la red. Al mismo tiempo, la industria busca la forma de movilizar y aplicar las enormes inversiones necesarias para el crecimiento futuro, por ejemplo para mejorar el acceso del sector residencial. Si Internet sufre un traspiés no será debido a la falta de tecnología, visión o motivación. Será debido a que no podemos hallar la dirección justa por la que marchar unidos hacia el futuro.

1.1.10. La evolución de Internet.

El crecimiento de Internet lo podemos valorar a través de los siguientes datos, elaborados a partir de los datos de (Hobbes, 2000) y basado también en los que ofrece (Gray, 1996).

El crecimiento del número de servidores es muy importante y con el nuevo mecanismo de investigación, aplicado a partir de Enero de 1998, que es mucho más preciso los datos aumentan con respecto al antiguo sistema de investigación.

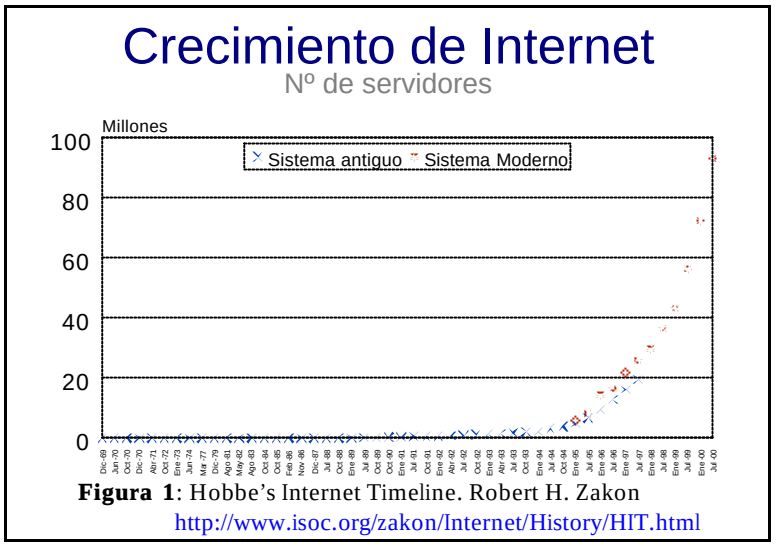


Figura 1: Hobbe's Internet Timeline. Robert H. Zakon
http://www.isoc.org/zakon/Internet/History/HIT.html

Respecto del número de redes el incremento es también muy acusado, ofreciendo una idea clara de la importancia y desarrollo de Internet.

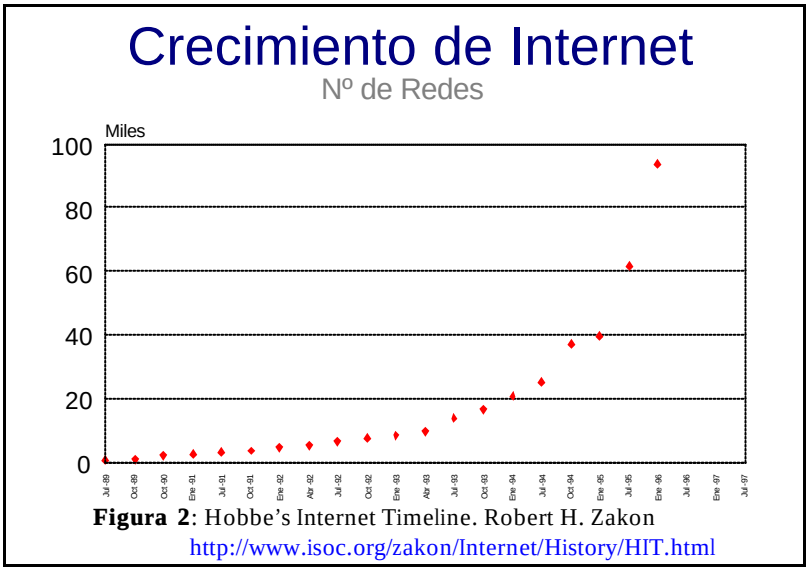


Figura 2: Hobbe's Internet Timeline. Robert H. Zakon
http://www.isoc.org/zakon/Internet/History/HIT.html

El incremento en el número de dominios también es espectacular.

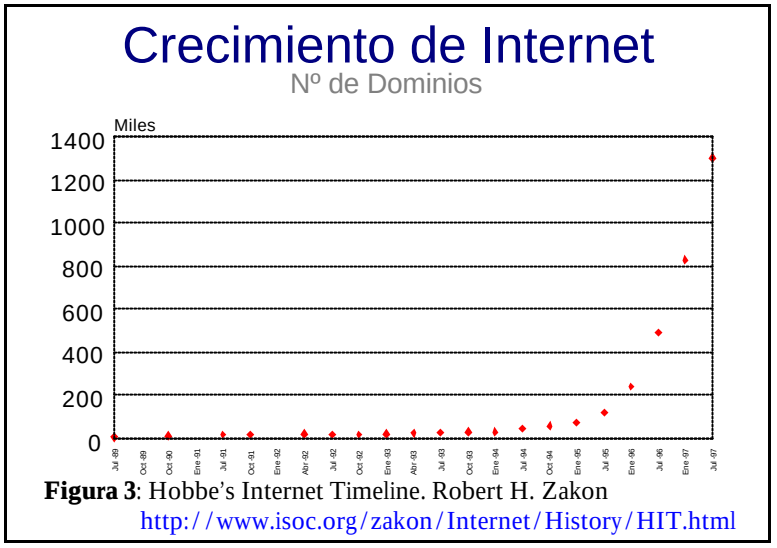


Figura 3: Hobbe's Internet Timeline. Robert H. Zakon
http://www.isoc.org/zakon/Internet/History/HIT.html

Estos datos nos ofrecen una idea de la importancia en el desarrollo de Internet y de la influencia que tiene, cada vez mayor, al convertirse en una tecnología ampliamente utilizada.

1.2. El World Wide Web.

La información sobre el World-Wide Web es amplia y extensa, aunque nosotros nos limitaremos a comentar algunos de los aspectos más interesantes, utilizando como base los trabajos de (Adell, 1994), (Adell, 1994b) realizaremos esta breve historia, que nos permitirá centrar los aspectos más interesantes, finalizando con unos datos de crecimiento que nos muestran la importancia del Web en la actualidad.

1.2.1. Introducción.

A finales de la década de los ochenta la interconexión de miles de redes de área local había convertido Internet en el mayor almacén de datos que jamás hubiese existido, pero también en el más caótico. Las posibilidades eran enormes, pero las dificultades resultaban frustrantes: formatos incompatibles, programas distintos, protocolos heterogéneos, etc. Se imponía pues la necesidad de simplificar

el acceso a este caudal de información, hacerlo más sencillo y homogéneo. WAIS, desarrollado a partir de 1989 por un grupo de empresas (Kahle, 1989) sólo fué una solución parcial: los datos debían indexarse con el nuevo software y distribuirse por medio de un nuevo protocolo, es decir, había que realizar un trabajo de adaptación de lo ya existente al nuevo sistema. El Gopher de la Universidad de Minnesota (Lindner, 1994), ampliamente difundido desde 1991, aportó algo más: por medio de un sistema simple de ventanas (o de menús) se accede a todo tipo de archivos de texto, imágenes, bases de datos, etc., sin tener que preocuparse por su localización física en la red, el formato o el protocólo de recuperación: ftp y wais, por ejemplo, son protocolos que el gopher maneja desde el principio, además del suyo propio. Un interface unificado para el acceso a información distribuida: este ha sido el objetivo del gopher, y también el del Web. El proyecto World-Wide Web del CERN (*Centre Européenne pour la Recherche Nucléaire)* ha venido a suponer otra vuelta de tuerca en el intento de poner efectivamente al alcance de los usuarios el espacio virtual de conocimiento que es Internet.

1.2.2. Hipertexto e hipermedia

La experiencia de la proliferación del conocimiento y de la angustia derivada de no poder abarcarlo todo no es nueva, no ha surgido con los ordernadores y la conectividad. Ya en 1945 Vannevar Bush (Bush, 1945) se lamentaba:

> La suma de la experiencia humana se está expandiendo a un ritmo prodigioso y los medios que utilizamos para seguir el hilo a través del consiguiente laberinto de ítems momentáneamente importantes son los mismos que usábamos en los días de los barcos de vela. (Bush, 1945).

En su opinión el problema no era tanto una cantidad excesiva de publicaciones como el nulo avance de las tecnologías con que se gestionaba su manejo. Con los rudimentos tecnológicos de su época en mente, Bush fue capaz de idear un sistema llamado memex que permitiría archivar el conocimiento de un modo más eficaz: una especie de escritorio futurista en el que se guardarían, microfilmados, los libros, actas, ficheros, etc. Cada elemento de información se

visualizaría en pantalla tecleando su código mnemotécnico correspondiente y, esto es lo más importante, podríamos registrar las conexiones observadas entre elementos distintos. Un usuario del memex que contase con una buena base de datos podría anotar conexiones entre, digamos, un artículo de enciclopedia sobre el escritor angloamericano H. Ph. Lovecraft, una fotografía suya y alguno de sus cuentos. Al leer el artículo, la simple pulsación de un boton le permitiría hojear "El horror de Dunwich" o visualizar la fotografía. Más tarde podría conectar con este conjunto la biografía de Lovecraft escrita por Pierre Bourbonnais.

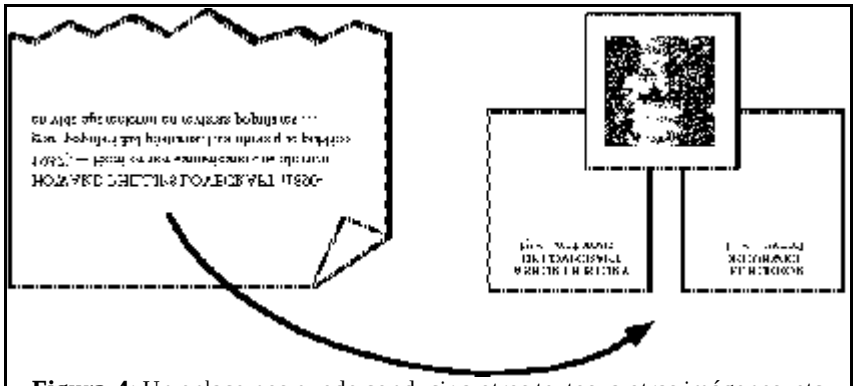


Figura 4: Un enlace nos puede conducir a otros textos, a otras imágenes, etc.

Bush remarcaba que este tipo de asociación no lineal de ideas era el modo de funcionamiento natural de la mente humana, y confiaba en que dispositivos semejantes al memex lo reproducirían en el futuro más adecuadamente. Es un hecho que los artículos de una enciclopedia, las notas al pie o las referencias bibliográficas contienen conexiones no lineales de aquel tipo, pero los medios tradicionales resultan inadecuados para gestionarlas. Cuando nos encontramos con una referencia bibliográfica que nos interesa, todo lo que podemos hacer es acudir a una biblioteca o una librería. Con el memex, idealmente, pulsaríamos un botón para consultar en nuestra pantalla el libro en cuestión. En el futuro, profetizaba Bush, las enciclopedias serían redes de conexiones que el usuario podría anotar y modificar a su antojo.

Bush era un visionario. En 1945 sus ideas no eran técnicamente realizables. Ni lo eran aún en 1965, cuando otro visionario, Ted Nelson, las ordenó conceptualmente. Fue Nelson quien acuñó el término `**hipertexto**' para referirse a "un cuerpo de material escrito o gráfico interconectado de un modo complejo que no se puede representar convenientemente sobre el papel; puede contener anotaciones, adiciones y notas de los estudiosos que lo examinan" (Nelson 1965). La idea es que el lector examina los nodos de una red, y pasa de unos a otros siguiendo las conexiones (links, en inglés). El hecho de que los nodos pueden contener texto, pero también pueden integrar otros medios: imagen, sonido, etc. es lo que se quiere remarcar con otro término complementario: `**hipermedia**'.

Durante las dos décadas siguientes se vivió el auge de los ordenadores, el almacenamiento digital y las redes. El propio Nelson cobró conciencia de lo apropiado de estas nuevas tecnologías para la realización del sueño de una red de elementos de información libremente accesible alrededor del mundo. Sin embargo, se diría que sus ideas sólo han llegado a concretarse recientemente con el World-Wide Web. Ha habido numerosos proyecto de sistemas hipertexto, encontrando una relación exhaustiva en (Balasubramaniam, 1994).

1.2.3. El proyecto World-Wide Web .

En 1989 la red mundial de datos, el memex global, ya existía en potencia. Internet, que se originó en el ámbito militar durante la guerra fría (Hardy 1993), se había desarrollado más allá de los propósitos originales como resultado de su uso por parte de la comunidad científica internacional, que necesitaba nuevos sistemas de distribución de la información. Lo único que se requería, como decíamos al principio del artículo, eran vías de acceso sencillas y homogéneas. Este era uno de los objetivos que Tim Berners-Lee se planteó en 1989 cuando presentó a sus superiores del CERN la propuesta original para el proyecto World-Wide Web. Otro era la posibilidad de gestionar conexiones no lineales.

`World Wide-Web' (abreviado `Web'; escrito también `WWW' o incluso `W3') significa algo así como `red (o telaraña) global'. La propaganda oficial del CERN lo define como un "sistema hipermedia distribuido" (Boutell 1994). En principió se pensó como un medio para la distribución de la información entre equipos de investigadores geográficamente dispersos; concretamente se dirigía a la

comunidad de físicos de altas energías vinculados al CERN. En su primera propuesta, Berners-Lee exponía las desventajas del uso de sistemas incompatibles e inconexos:

> "En el CERN, una diversidad de datos está ya disponible: informes, datos experimentales, datos personales, listas de direcciones de correo electrónico, documentación informática, documentación experimental y muchos otros conjuntos de datos están girando contínuamente en discos de ordenadores. Es sin embargo imposible 'saltar' de un conjunto a otro de una manera automática: una vez has encontrado que el nombre de Joe Bloggs se lista en una descripción incompleta de algun software en línea, no se encuentra directamente su dirección actual de correo electrónico. Usualmente, tendrás que utilizar un método de consulta distinto en un ordenador distinto con un interface distinto. Una vez has localizado la información, es difícil guardar sus conexiones o hacer una anotación privada que puedas después encontrar rápidamente." (Berners-Lee, 1994)

La conclusión era que "hay un enorme beneficio potencial en la integración de una variedad de sistemas de un modo que permita a los usuarios seguir conexiones que apuntan de un elemento de información a otro".

Se pretendía pues que los recursos disponibles en formato electrónico, que residen en ordenadores distintos conectados a la red, fuesen accesibles para cada investigador desde su terminal, de un modo transparente y exento de dificultades, sin necesidad de aprender a utilizar varios programas distintos. Además, debería posibilitarse el salto entre elementos de información conexos. Los recursos existentes deberían integrarse en una red hipertextual distribuida gestionada por ordenadores.

Las primeras instalaciones del WWW para uso interno del CERN estuvieron listas en 1991. Ese mismo año el sistema se abrió ya a Internet. Desde entonces, para acceder al World-Wide Web no se requiere más que un terminal VT conectado a Internet, pero la máxima facilidad de uso y el máximo rendimiento se alcanzan

con una pantalla gráfica (un modelo Next o Macintosh, un X-Terminal o un PC con tarjeta gráfica). Entonces el sistema nos ofrece hipertextos como el que muestra la figura 2, nodos de la telaraña global. Las palabras subrayadas, y las imágenes recuadradas, son links que nos conducen a otros nodos. Para viajar hasta ellos basta con situarse con el ratón sobre el link y pulsar el botón. El nodo de llegada puede ser otro hipertexto, o también un nodo no hipertextual integrado en la red: un servidor gopher, un grupo de netnews, una búsqueda en una base de datos WAIS, etc.

El éxito del WWW, el crecimiento de la telaraña, ha sido espectacular. Durante 1993 se pasó de 50 a 500 nodos. En 1994 se contabilizan ya miles de servidores en el WWW que distribuyen todo tipo de información (de ellos, trece en España; el primero fue el del Departamento de Educación de la Universitat Jaume I, en septiembre de 1993).

1.2.4. La arquitectura del World-Wide Web.

El diseño del World-Wide Web sigue el modelo cliente-servidor: un paradigma de división del trabajo informático en el que las tareas se reparten entre un número de clientes que efectúan peticiones de servicios de acuerdo con un protocolo, y un número de servidores que las atienden. En el Web, nuestras estaciones de trabajo son clientes que demandan hipertextos a los servidores. Para poner en marcha un sistema como como éste ha sido necesario:

a) Diseñar e implementar un nuevo protocolo que permitiera realizar saltos hipertextuales, esto es, de un nodo o lexia de origen a uno de destino, que podria ser un texto o parte de un texto, una imagen, un sonido, una animación, fragmento de vídeo, etc. Es decir, cualquier tipo de información en formato electrónico. Este protololo se denomina HTTP (HyperText Transfer Protocol) y es el "lenguaje" que "hablan" los servidores del WWW.

b) Inventar un lenguaje para representar hipertextos que incluyera información sobre la estructura y el formato de representación y, especialmente, indicar origen y destino de saltos hipertextuales. Este lenguaje es el HTML o (HyperTextex markup Language).

c) Idear una forma de codificar las instrucciones para los saltos hipertextuales de un objeto a otro de Internet. Dada la variedad de protocolos, y por tanto, formas de almacenamiento y recuperación de la información, en uso en Internet, esta información es vital para que los clientes (ver el siguiente punto) puedan acceder a dicha información.

d) Desarrollar aplicaciones cliente para todo tipo de plataforma y resolver el problema de cómo acceder a información que está almacenada y es accesible a través de protocolos diversos (FTP, NNTP, Gopher, HTTP, X.500, WAIS, etc.) y representar información multiformato (texto, gráficos, sonidos, fragmentos de vídeo, etc.). A este fin se han desarrollado diversos clientes, entre los que destaca la familia Mosaic, del NCSA (National Center for Supercomputer Applications) de la Universidad de Chicago, y su sucesor Netscape Navigator, de Netscape Communications Corporation.

Pero, veamos con cierto detenimiento los rasgos más sobresalientes de estos elementos clave del sistema.

1.2.4.1. HTTP: HyperText Transfer Protocol .

El HTTP (HyperText Transfer Protocol) es el protocolo de alto nivel del World-Wide Web que rige el intercambio de mensajes entre clientes y servidores del Web. Un protocolo es:

> "Una descripción formal de los formatos de los mensajes y las reglas que deben seguir dos ordenadores para intercambiar dichos mensajes. Los protocolos pueden describir detalles de bajo nivel de los interfaces de máquina a máquina (por ejemplo, el orden en el cual deben enviarse bits y bytes a través de un cable) o intercambios de alto nivel entre programas (por ejemplo, la forma en que dos programas transfieren un fichero a través de Internet)." (Malkin, 1993).

El HTTP es un protocolo genérico orientado a objetos que no mantiene la conexión entre transacciones (Berners-Lee, 1993b). Ha sido especialmente

diseñado para atender las exigencias de un sistema hipermedia distribuido como es el World-Wide Web. Sus características principales son:

Ligereza: reduce la comunicación entre clientes y servidores a intercambios discretos, de modo que no sobrecarga la red y permite saltos hipertextuales rápidos.

Generalidad: puede utilizarse para transferir cualquier tipo de datos, según el estándar MIME (Multipurpose Internet Mail Extensions, es una norma estándard de Internet para la transmisión de objetos multimedia que aparece en RFC 1521 y RFC 1522). Esto incluye también los que desarrollen en el futuro, ya que el cliente y el servidor pueden negociar en cualquier momento el modo de representación de los datos: el cliente notifica al servidor una lista de formatos que entiende, y en adelante el servidor sólo remitirá al cliente datos que este sea capaz de manejar. El cliente debe aceptar al menos dos formatos: text/plain (texto normal) y text/html (hipertexto codificado en HTML: el lenguaje en el que se escriben los hipertextos del Web --véase el apartado siguiente).

Extensibilidad: contempla distintos tipos de transacción entre clientes y servidores ("métodos", en la jerga HTTP), y la futura implementación de otros nuevos. Esto abre posibilidades más allá de la simple recuperación de objetos de la red: búsquedas, anotaciones, etc.

El esquema básico de cualquier transacción HTTP entre un cliente y un servidor es el siguiente (Berners-Lee, 1993):

Conexión:

El cliente establece una conexión con el servidor a través del puerto 80 (puerto estándar), u otro especificado.

Petición:

El cliente envía una petición al servidor.

Respuesta:

El servidor envía al cliente la respuesta (esto es, el objeto demandado o un código de error).

Cierre:

Ambas partes cierran la conexión.

La eficiencia del HTTP posibilita la transmisión de objetos multimedia y la realización de saltos hipertextuales con una rapidez razonable.

1.2.4.2. HTML: HyperText Markup Language.

El HTML (HyperText Markup Language) es el lenguaje en el que se escriben los hipertextos del World-Wide Web. Cumple la norma SGML (Standard Generalized Markup Language, que es la norma ISO 8879:1986)), y permite añadir a un documento de texto:

- La especificación de estructuras del texto. Por ejemplo, títulos, encabezamientos, límites de los párrafos, listas de elementos.
- Estilos: texto enfatizado, citas, etc.
- Objetos multimedia: imágenes o sonido, pongamos por caso.
- Conexiones hipertextuales a otros objetos de la red: partes sensibles del documento desde dónde podríamos saltar otras partes del Web.

Todo este "valor añadido" al texto se codifica como etiquetas ("tags", en la jerga) que se insertan en el propio texto. Un ejemplo lo podemos ver en la Figura 4.

Las etiquetas del HTML se delimitan por medio de los signos < y >. Por ejemplo, la etiqueta <P> marca el inicio de cada párrafo. Otras, la mayor parte, van por parejas: <TITLE> y </TITLE> abren y cierran, respectivamente, el título del documento.

Los links se abren y cierran con las etiquetas <A> y </A>. El objeto de la red a donde nos lleva el link se codifica en la etiqueta de apertura por medio de una notación que se ha convertido de hecho en un estándar de Internet: los llamados URL.

Figura 5: Ejemplo de documento HTML.

1.2.4.3. URL: Uniform Resource Locator.

Los URL (Uniform Resource Locator) son una notación estándar para la especificación de recursos presentes en Internet. Constituyen la piedra angular del Web, ya que hacen posible que un link de HTML se refiera a cualquier objeto de la red.

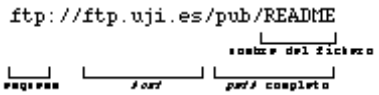
Un URL representa de un modo compacto la localización y el método de acceso de cualquier recurso de la red (Berners-Lee, 1994b), (Weider, 1994). No sólo hay más de dos millones de ordenadores conectados a los varios miles de redes que forman Internet, sino que existen múltiples protocolos o formas diferentes de acceder a la información (ftp, gopher, http, etc.). Los URL aportan esos dos datos esenciales: dónde se encuentra un recurso y cómo se puede acceder a él.

La sintaxis de los URL es la siguiente:

URL: < esquema > : < parte-específica-del-esquema >

El esquema es un término convenido que representa el método de acceso a un recurso. La parte específica del esquema informa sobre su localización en la red, de un modo que depende de cada método de acceso. Un ejemplo nos ayudará a entender esto.

Cuando utilizamos ftp anónimo para copiar un fichero de un ordenador remoto a nuestro ordenador necesitamos saber lo siguiente: host o nombre del ordenador remoto donde se encuentra el fichero y path o via que conduce al fichero dentro de la estructura de ficheros del ordenador remoto. Supongamos que el fichero se llama README, y que está en el directorio pub del host ftp.uji.es; el URL de tal objeto sería éste:



Al recuperar un fichero mediante ftp anónimo usamos "anonymous" como nombre de usuario, y nuestra dirección de correo electrónico como password. En los URL esta información se omite dado que es conocida. Sin embargo, es posible incluirla si, por ejemplo, no se trata de ftp anónimo, sino que se necesita especificar un usuario real y su password. La sintaxis genérica de los URL para objetos accesibles por ftp es la siguiente:

URL:ftp://[user[:password]@]host[:port]/path[;type = < typdecode >]

El "port" puede omitirse si el servidor de ftp emplea el port estándar de ftp (el 21). Este principio de omitir lo ya conocido se sigue en todos los URL. Si los

distintos servidores siguen las recomendaciones de Internet no es necesario incluir información redundante.

El "path" es la lista ordenada de subdirectorios por los que hay que pasar para llegar al fichero, separados por "/", seguida del nombre del fichero.

El "type" es "d", "a", "i". "d" indica que se requiere la transmisión de una lista de nombres de ficheros (un directorio). "a" solicita una transmisión de líneas de texto. "i" solicita una transmisión binaria.

En la actualidad existen esquemas definidos para los siguientes servicios:

Esquema	**Sintaxis**
ftp (File Transfer Protocol)	ftp://user:password@host:port/path;type = <typecode>
http (HyperText Transfer Protocol)	http://<host>:<port>/<path>?<searchpart>
gopher (gopher)	gopher://<host>:<port>/<gopher-path>
mailto (correo electrónico)	mailto:<rfc822-addr-spec>
news (USENET news)	news:<newsgroup-name>
nntp (USENET news especificando unservidor nntp, NetNews Transfer Protocol	nntp://<host>:<port>/<newsgroup-name>/<article-number >
wais (Wide Area Information Server)	wais://<host>:<port>/<database> o wais://<host>:<port>/<database>?<search> o wais://<host>:<port>/<database>/<wtype>/<wpath>

Ejemplos

ftp://milano.usal.es/Software/fichero.doc

http://www.usal.es

(URL de la página de entrada del servidor Web de la Universidad de Salamanca)

gopher://gopher.uji.es

(URL de la entrada del servidor gopher del "Servei d'Informació del Campus (sic) de la Universitat Jaume I")

mailto:berrocal@gugu.usal.es

(Este URL posibilita el envío de un mensaje de correo electrónico)

news:comp.infosystems.gopher

(URL del grupo de news comp.infosystems.gopher)

nntp://news.uji.es/comp.infosystems.gopher

(Este URL especifica el grupo de news comp.infosystems.gopher almacenado en el servidor news.uji.es)

wais://wais.uji.es/tractatus?ethics

(Este URL especifica la búsqueda del término "ethics" en la base de datos "tractatus" del servidor WAIS wais.uji.es)

La utilidad, y la necesidad, de una notación que, como ésta, introduzca algo de orden en el caos de la red es obvia. Los URL se idearon para un proyecto concreto y limitado, el del WWW, pero ha cundido el ejemplo. Ahora mismo se está produciendo un amplio debate en el seno de Internet, concretado en un grupo de trabajo de la IETF (Internet Engineering Task Force) para el desarrollo de sistemas universales de designación y caracterización de objetos persistentes de la red, inspirados en los URL pero que irían más allá: debería ser posible, por ejemplo, asignar un URN (Uniform Resource Name) invariable para un objeto, aunque cambiara su path e incluso su método de acceso (Weider, 1994). Un sistema distribuido (similar al DNS o Domain Name System) resolveria un URN en uno o varios URL aplicando criterios de optimización de recursos (como proximidad al solicitante).

1.2.4.4. Internet como telaraña: El World Wide Web.

Al principio del artículo hemos presentado el World-Wide Web como un proyecto de integración de recursos de la red. Después hemos dicho que con el Web se cumplían los viejos anhelos de Vannevar Bush. Nos parece obvio que un

sistema de acceso a Internet debe ser hipermedia, porque la información no se ceñirá a un solo medio y porque ha de ser posible seguir las conexiones entre los elementos. Pero, ¿es el Web el sistema hipermedia perfecto? Es obvio que en algunos aspectos fundamentales dista de serlo. Sobre todo, en la posibilidad de personalizar las conexiones entre los elementos de información. Realizar anotaciones para comentar las páginas no es suficiente, y una adecuada personalización de páginas exige que escribamos nuestro propio código HTML, lo que no está (¿aún?) al alcance de todos.

El HTML, por su parte, también tiene puntos flacos. El hecho de que las marcas se integren en el propio texto dificulta el mantenimiento de éste. La modificación del texto hace necesario volver a aplicar las marcas.

También es arduo mantener los links, pero esto no es tanto un problema del HTML como del sistema de URL. Ya hemos dicho que se intenta superar los URL mediante la especificación de URN: nombres permanentes de objetos, independientes de sus localizaciones y métodos de acceso transitorios, que unos servidores de nombres resolverían en los URL correspondientes.

Finalmente, parece que no basta con el acceso hipermedia a la red. Internet continua siendo un almacen caótico. Sólo hemos ordenado el interface de usuario, el acceso a los datos, pero estos continuan desordenados. Para solventar este desorden se requieren sistemas de indexación y catalogación que pueden estar basados en los actuales, como WAIS.

La importancia del Web como sistema de difusión de la información se ve avalada por el importantísimo crecimiento tanto en número de sedes como en el número de páginas Web, que poseen un crecimiento claramente exponencial como además (Huberman, 1999) refleja en la recogida de datos sobre los motores Alexa e Infoseek. Trabajos que nos permiten valorar el crecimiento del Web son los de (Gray, 1995), (Gray, 1996), (Bray, 1996), (Coffman, 1998), (OCLC, 1999), (MIDS, 1999), (ISC, 2000), (Hobbes, 2000).

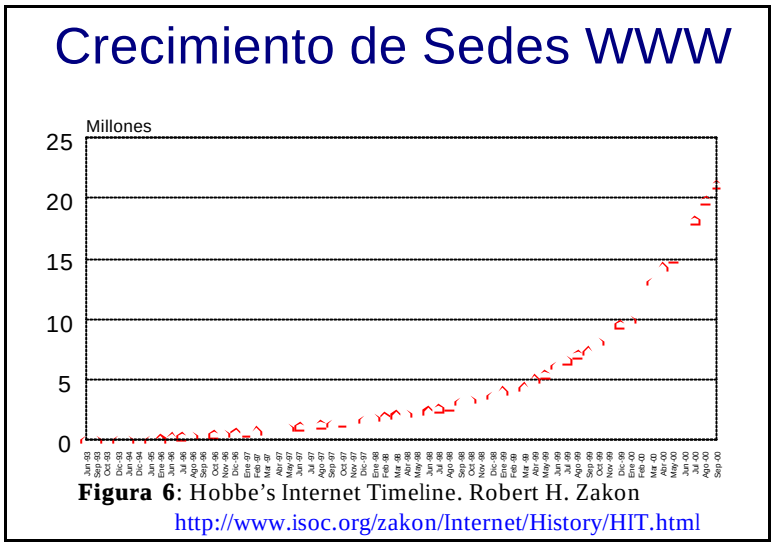


Figura 6: Hobbe's Internet Timeline. Robert H. Zakon
http://www.isoc.org/zakon/Internet/History/HIT.html

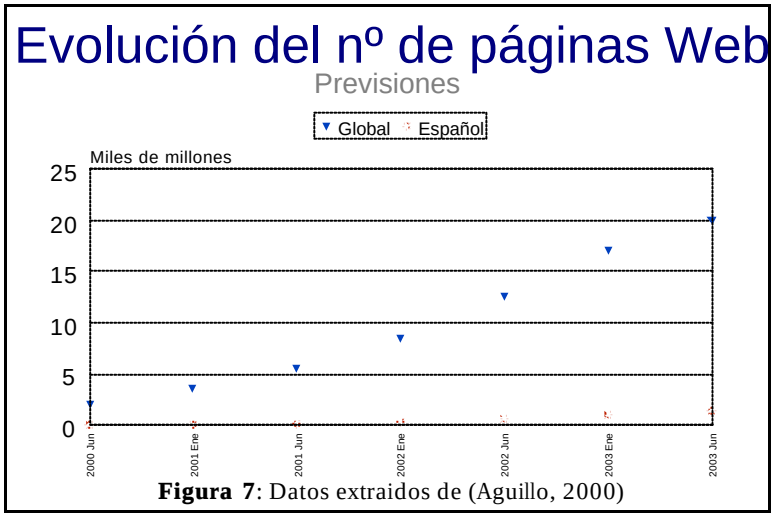


Figura 7: Datos extraidos de (Aguillo, 2000)

1.3. Historia de la bibliometría.

La historia de la bibliometría, amplia y plagada de hitos la ofrecemos aquí de forma muy abreviada, centrándonos en aquellos aspectos que nos permiten clarificar los aspectos de nuestro interés. Esta historia se ha realizado a partir de los

trabajos de (Osareh, 1996) y (Okubo, 1977) centrándonos en algunos de los episodios más relevantes de dicha historia.

Históricamente, la bibliometría tuvo su origen en occidente, y derivó de los estudios estadísticos de bibliografías (Egghe, 1990). Aunque el término bibliometría es bastante reciente (Pritchard, 1969) su empleo y utilización pueden verese desde 1890. (Campbell, 1896) trabajó con métodos estadísticos para estudiar las materias esparcidas en las publicaciones siendo seguramente el primer intento de estudio bibliométrico (Sengupta, 1992).

(Cole, 1917) estudió estadísticamente el crecimiento de la literatura en anatomía comparada durante 1550-1860, a través de las citas bibliográficas. (Hulme, 1923) empleó el término bibliografía estadística para describir cómo los métodos de la historia de la ciencia y de la tecnología pueden ser más comprensibles contando documentos (Garfield, 1977). Igualmente, Ranganathan creía que el análisis estadístico y matemático eran herramientas clave para todos los desarrollos y estudios de previsiones. Él formalmente sugirió el término *librametry* [sic] en 1948 e indicó que desde la aplicación de la estadística y las matemáticas han surgido nuevas especialidades tales como la biometría, econometría, psicometría, etc. y en los bibliotecarios deberían emplear adecuadamente las técnicas matemáticas y estadísticas para desarrollar la *librametry* [sic] (Sengupta, 1992).

En 1969, en el Seminario Anual para la Investigación de la Documentación y Centro de Entrenamiento (DRTC) Ranganathan claramente demostró la aplicación de las técnicas Librométricas.

(Pritchard, 1969) sustituyó el termino bibliografía estadística por el de bibliometría, mientras Nalimov y Mulchenko (Nalimov, 1969) usaban el término cienciometría. Numerosos autores aceptaron el término de Pritchard y le concedieron la paternidad de la palabra bibliometría. Sin embargo, (Fonseca , 1973) indica que el equivalente francés del término aparece en el trabajo de (Otlet, 1934) “Traite de documentation. Le livre sur le livre. Theorie et pratique”.

1.3.1. Definición de bibliometría.

Los principios de la medición de la productividad científica se desarrollan desde los años 60 de la mano de la bibliometría y podemos considerar a Derek John de Solla Price como su fundador (Polanco, 1995) y uno de sus más aventajados teóricos.

La palabra bibliometría tiene dos raíces: Biblio y Metría. Biblio deriva de la palabra griega y latina biblion, significando libro. El término metría, indica la ciencia de medir y deriva de la palabra greco-latina metricus o metrikos significando medida (Sengupta, 1992).

Como indica (López, 1986) el empleo de las matemáticas no es nuevo en el estudio de la producción científica. Algunos autores proponen la fecha de 1885 con los trabajos de AlPhonse de Condolle (Méndez, 1986), como el punto de partida de las técnicas bibliométricas. En nuestro país será Ortega y Gasset quien hable de la necesidad de hacer una estadística de las ideas (Ortega, 1935).

Según Pritchard, que acuñó el término (Pritchard, 1969), la bibliometría sería "la aplicación de métodos matemáticos y estadísticos a los libros y otros medios de comunicación". Parafraseando a Pritchard, (Fairthorne,1969) la define como "el tratamiento cuantitativo de las propiedades del discurso registrado y comportamiento de lo publicado". Posteriormente en 1976 la British Standards Institution describió la bibliometría como la aplicación de métodos matemáticos y estadísticos en el estudio del uso de los documentos y de los patrones de publicación (British, 1976).

(Hawkins, 1977) definió bibliometría de forma similar, pero en un lenguaje más simple, como la aplicación del análisis cuantitativo a las referencias bibliográficas de la literatura. (Lancaster, 1977) la describió como el estudio del empleo de patrones de escritores, de publicaciones y de literatura por la aplicación de diversos análisis estadísticos.

(Schrader, 1981) como profesor de bibliometría la definió más explícitamente como "el estudio científico del discurso registrado". (Potter, 1981) afirma que la bibliometría es un medio para el estudio y medida de todas las formas de comunicación escrita, sus autores y sus patrones de publicación.

(White, 1989) afirma que la bibliometría es "el estudio cuantitativo de la literatura según se refleja en su bibliografía". (Diodato, 1994) describe la bibliometría como el estudio de los patrones de publicación y comunicación en la distribución de información mediante el empleo de técnicas matemáticas y estadísticas.

Martínez de Sousa la define como la "técnica de investigación bibliológica que tiene por fin, por un lado, analizar el tamaño, crecimiento y distribución de la bibliografía en un campo determinado, y por otro, estudiar la estructura social de los grupos que la producen y la utilizan" (Martínez, 1989)

Por otra parte, (Braun, 1985) coincide con la definición de cienciometría de (Nalimov, 1969) diciendo que estos métodos cuantitativos relacionados con el análisis de la ciencia, vistos como un proceso de información, se refieren a la cienciometría. Bookstein define cienciometría claramente como "la ciencia de medir la ciencia". Braun, sin embargo, recalcó la importancia de distinguir entre bibliometría y cienciometría según el tema y propósito del término; sin embargo, sus métodos son muy similares y en ocasiones idénticos.

El propósito principal de la bibliometría es mejorar la documentación científica, las actividades de información y comunicación mediante el análisis cuantitativo de las colecciones bibliotecarias y de sus servicios. Mientras las técnicas cienciométricas se emplean, para contribuir a una mejor comprensión del mecanismo de investigación científica como actividad social, mediante el análisis cuantitativo de la generación, propagación y utilización de diferentes aspectos de la información científica.

(Ravichandra, 1993) sugiere que la informetría se emplea desde mediados de la década de los 80 y su alcance cubre la bibliometría y la cienciometría, así como otros estudios cuantitativos relacionados con las ciencias de la información. "Informetría connota el empleo y desarrollo de una gran variedad de medidas para el estudio y análisis de varias propiedades de la información en general y de los documentos en particular".

Xavier Polanco precisa más y considera otros términos como la cienciometría, que "se puede considerar como la bibliometría especializada en el

dominio de la información científica y técnica" y la infometría "término adoptado en 1987 por la FID -Federación Internacional de Documentación- para designar el conjunto de actividades de medición relativas a la información, abarcando tanto la bibliometría como la cienciometría" (Polanco, 1995)

1.3.2. Alcance de la bibliometría.

(Pritchard, 1969) indica que los objetivos de la bibliometría son "arrojar luz sobre el proceso de la comunicación escrita y de la naturaleza y dirección del desarrollo de una disciplina, mediante la cuenta y análisis de varias características de la comunicación escrita.

(Nicholas, 1978) dividió los estudios bibliométricos en dos amplios grupos: estudios descriptivos (tratando con características de la literatura) y estudios de comportamiento, en ocasiones referido a estudios de citas, pero no restringido a ellos (tratando con las relaciones generadas entre los componentes de la literatura).

Por otra parte (Stevens, 1953) dividió los estudios bibliométricos en dos áreas básicas y varias subáreas.

1. Cuentas de la productividad o área descriptiva:
    a. Países (localizaciones geográficas)
    b. Diferentes periodos de tiempo
    c. Diferentes disciplinas (campos de materia)
2. Cuentas sobre el uso de la literatura o área evaluativa:
    a. Referencia
    b. Citas

(Diodato, 1994) describe tres áreas en la investigación bibliométrica:

1. Leyes bibliométricas o distribuciones, tales como las leyes de Bradford, Lotka y Zipf

2. Análisis de citas

3. Indicadores de rendimiento de la investigación.

(Callon, 1995) agrupa los análisis bibliométricos en función del tipo de resultados que nos proporcionan en:

1. Indicadores de actividad: cuando tratamos de cuantificar el impacto y actividad de los investigadores.

2. Indicadores de relación: que persigue mostrar las relaciones e interacciones entre investigadores y los distintos campos, ver su contenido y evolución.

Antes de pasar a ver el término cibermetría, merece la pena indicar, que ya algun autor (Turnbull, 1996) hablaba de bibliometría y World-Wide Web e indicaba que al Web se pueden aplicar las mismas técnicas bibliométricas que a los artículos, así como el análisis de citas y de cocitas. Sin embargo no indica la metodología necesaria para hacerlo y finalmente se centra en el estudio de los ficheros Log, como aquello que se puede estudiar, olvidándose de la estructura hipertexto y del valor de los enlaces.

También (Larson, 1996) habla de bibliometría y del World Wide Web, de forma que la aplicación de las técnicas bibliométricas al Web nos permiten analizar el ciberespacio.

Algún otro autor (Bossy, 1995) hace referencia a la redmetría como sistema de estudio de la red Internet, pero lo cierto es que en su trabajo no aporta nada, excepto que deben aplicarse técnicas similares a las bibliométricas que nos permitan conocer el Web.

1.4. Cibermetría.

Para aclarar el término nada mejor que trabajar con las ideas del autor (Shiri, 1998) que acuñó el término y ofrecer sus ideas sobre el mismo. Disponemos

así de información de primera mano de lo que este autor considera que la cibermetría debe estudiar y como la considera.

Ningún periodo de la historia ha presenciado cambios y desafíos tan profundos en la organización y la difusión de la información como el actual. Los descubrimientos a nivel mundial cada vez mayores en tecnologías de la información y en servicios nos han incitado a dar pasos en un mundo artificial cuyos elementos, fenómenos y seres son totalmente diferentes del mundo en el cual vivimos. El mundo que se conoce como " Ciberespacio ". Un espacio en el que la principal criatura viva es la información. Alvin Toffler lo denominó como Infoesfera.

Quizás, cuando William Gibson acuñó el término Ciberespacio (Gibson, 1984) en su Neuromancer no podría imaginarse cómo los horizontes insondables tendrían un futuro tan cercano. Hoy, estamos haciendo frente a un Ciberespacio mucho más complejo y multidimensional que el percibido por Gibson. El término Ciberespacio, ampliamente utilizado hoy en día, define las complejas comunicaciones del Web a nivel mundial(Haynes, 1995). Hoy, hay, no solamente los profesionales de la información, los documentalistas y los bibliotecarios que utilizan los amplios potenciales del Ciberespacio, sino también todas las personas de una gran variedad de profesiones y de empresas que utilizan las diferentes capacidades de este espacio de la información. Esta es la razón por la cual hay varias opiniones respecto a la noción de Ciberespacio. Otra definición de Ciberespacio la considera como un espacio de posibilidades de computación interactivas, donde están disponibles los ordenadores y su contenido para los usuarios de cualquier ordenador dondequiera que se encuentren (Bauwens, 1996). Esta definición continúa con una interpretación orientada hacia la información, que considera que el Ciberespacio es donde se almacenan y se transmiten cada vez más información y conocimiento, siendo muy importante el lugar donde estamos al comunicarnos con un colega a través de los ordenadores. Haynes también cree que el Ciberespacio es más amplio que el World Wide Web e incluso que Internet. Hay varios miles de redes de comunicaciones que demuestran la globalidad del ciberespacio (Haynes, 1995).

Hojeando estas definiciones, un aspecto que sobresale en la mayoría de ellas, es que inciden más en los medios que en el significado. La explicación de

Bauwens Ciberespacio parece ser más intensa en el aspecto de la información que las otras definiciones. Acentuando la importancia de la información en nuestra discusión de Ciberespacio, quisiera señalar la atención de los profesionales de la información a esta realidad indiscutible de que la información es el corazón de tal atmósfera, aunque en la creación de dicho espacio intervengan muchas clases de tecnologías del ordenador, de la telecomunicación y de la información. Así pues, esta ciberinformacion marca una nueva frontera de la investigación de la información.

Generalmente hablando, la ciberinformación implica la información comunicada a través de medios electrónicos. Para clarificar más el concepto de ciberinformación, parece necesario indicar los medios de información principales, que constituyen la base del Ciberespacio. Pueden ser detallados como sigue:

1. Redes de información de todas las clases y alcances

2. Bases de datos y metabases en línea

3. Herramientas de Internet y medios incluyendo homepage, sedes Web, E-mail, grupos de discusión y de noticias

4. Escuelas virtuales, universidades y organizaciones

5. Sistemas del tablón de anuncios

6. Conferencias electrónicas, asociaciones, y sociedades

7. Libros electrónicos, bibliotecas, archivos y servicios información

8. Sistemas de información multimedia, hipermedia, polymedia y telemedia

Es evidente que la lista anterior no está completa, pero proporciona algunos componentes importantes del Ciberespacio. Muchos otros términos se pueden agregar a la lista anterior precedidos de adjetivos como "electrónico " o "digital" y el prefijo "Cibe ".

Un aspecto importante, que debería ser recalcado, es que muchas actividades, individuales y sociales, que tienen algún tipo de comunicación y de intercambio de información se incorporan cada vez más al ciberuniverso.

1.4.1. Los acercamientos cuantitativos anteriores a la ciberinformación.

El crecimiento rápido y cada vez mayor en la información electrónica junto con los amplios potenciales de las tecnologías y de los medios de información recientemente emergentes, han atraído la atención de los investigadores de la información para reflejar sobre la medida y la métrica cuantitativas de las fuentes de información, de los servicios y de los medios en esta esfera emergente, el cibercosmos. Las investigaciones principales en este área se han emprendido desde 1996 hacia adelante. Arnzen se refiere a cibercitas y a los ejemplos de citas del correo electrónico, website, ftp, Gopher, USENET, o listas de correo (Arnzen, 1996).

Clausen ha realizado otra investigación. Usando métodos de investigación mediante el empleo de encuestas, ha estudiado el uso de los recursos de Internet, usuarios y sus categorías de edad, el número de usuarios y conferencias electrónicas. En Dinamarca, se ha esforzado en cuantificar los hábitos de los usuarios de Internet y también sus actitudes hacia Internet como recurso de la información (Clausen, 1996).

Un estudio cuantitativo orientado a los medios, referente a Internet, es el trabajo de (McMurdo, 1996). Su trabajo gira principalmente sobre los medios, más que sobre la información. Contando los host de Internet y sus dominios, la distribución de host por dominios, crecimiento del Web, número de hosts de las sedes Web y las relaciones de transformación de las sedes Web están entre los parámetros principales que él ha estudiado a través de su investigación. En el estudio se han utilizado algunas fuentes de información estadística y demográfica sobre internet.(McMurdo, 1996).

Una de las investigaciones principales realizadas sobre la métrica del cibermedio es la de Almind y de Ingwersen (Almind, 1997). Procuraron introducir la aplicación de métodos informétricos al World Wide Web (WWW) denominándolo " Webmetría ". Realizaron un estudio comparando la proporción Danesa de WWW a la de otros países nórdicos. La metodología usada era de análisis bibliométrico. Dentro de este estudio se analizaron cinco aspectos fundamentalmente:

- Un análisis de la posición de Dinamarca respecto al Web

- Un análisis de la distribución de las paginas Web Danesas en grandes centros de enseñanza en Dinamarca

- Un análisis de la distribución de los dominios científicos sobre una muestra

- Un análisis de la distribución de las paginas Web sobre el tipo de documento

- Un análisis de la distribución de frecuencias seleccionadas para una muestra de páginas Web

Este estudio también ha explorado el número medio de hiperenlaces por página Web y la densidad de enlaces para los diferentes tipos de dominio (Almind, 1997). Este estudio webométrico, era una investigación de todas las comunicaciones basadas en la red usando la informetría u otras medidas cuantitativas. Sin embargo, debemos considerar que se han centrado principalmente en el análisis cuantitativo del World Wide Web.

También el trabajo de (Abraham, 1997) y empleando el término Webmetría habla de la necesidad de aplicar técnicas de redes neuronales para el mejor conocimiento del Web, representando las conexiones de los nodos mediante número reales, que indicaran la fuerza de la conexión. Indica la necesidad de emplear matrices, aunque en ningún momento hace referencia a la teoría de grafos.

En 1997 una investigación, en la Escuela Real de Bibliotecarios de Dinamarca, dirigido a explorar por estudios cuantitativos ciertos fenómenos y acontecimientos actuales de la información. Uno de los primeros objetivos de esta investigación es el análisis de la creación, uso y del estudio de las homepages Danesas/Nórdicas. Este estudio también se ha referido a Internetmetría (Informetría). Parece estar más orientado a la información que las investigaciones anteriores. Una investigación reciente se ha realizado sobre el factor de impacto del Web (Ingwersen, 1998). Este estudio informa sobre las investigaciones para ver la viabilidad y la fiabilidad en el cálculo del factor de impacto de las sedes Web llamado factor de impacto del Web. El estudio demuestra que el factor de impacto del Web es calculable y fiable con la precaución necesaria para estimar el número de las paginas del Web que señalan a las paginas de una sede determinada.

(Dahal, 1999) aplicó las leyes bibliométricas al análisis del desarrollo de los sistemas de información en ciencia y tecnología del Nepal, empleando finalmente el término cibermetría para explicar las técnicas empleadas.

Parece evidente que la aplicación de la métrica y de las medidas cuantitativas a la información electrónica se está convirtiendo cada vez más un área significativa para la investigación.

1.4.2. Cibermetría: Otra dimensión en la investigación de la información.

El incremento en la transición de los materiales impresos a los recursos electrónicos y a recursos de red ha originado a su alrededor nuevas perspectivas para estudiar las fuentes, los servicios y los medios de información. Es decir, si queremos estar enterados de qué información aparece en nuestro entorno, el análisis cuantitativo y el estudio de los fenómenos que operan dentro de este entorno es tan importante como ha sido el estudio cuantitativo de las características de los materiales impresos en el pasado. A través de estos estudios podemos hacer una estimación de qué se conoce como información electrónica y evaluar las características de tal información.

¿Qué significa el término cibermetría? Por este término debemos entender la medida, el estudio, y el análisis cuantitativo de todas las clases de información y de los medios de información que existen y que funcionan dentro del ciberespacio empleando las técnicas bibliométricas, cienciométricas e informétricas.

El principal incentivo de la cibermetría es la amplia variedad de nuevos medios electrónicos por medio de los cuales se comunica una amplísima gama de informaciones. Desde que los servicios de información tradicional y las fuentes, en gran parte, han sido transformadas en nuevos soportes y formatos que reclaman un cambio en el acercamiento a los estudios de la información, la necesidad urgente de reconsiderar nuestros esfuerzos investigadores en esta área parecen evidentes.

Las redes de información como mecanismo importante para la comunicación de la información puede considerarse como una de las áreas principales para ser estudiada. Existen redes funcionando a nivel nacional, internacional o globalmente. El número de cada clase de red, su cobertura temática, el número de usuarios y su dispersión geográfica son elementos para su

investigación. Internet como red de información global nos ha provisto de una amplia gama de servicios informativos y de medios. Las sedes Web, las homepages, el E-mail, grupos de discusión y de noticias son algunas de las herramientas principales de Internet a través de las cuales todas las clases de información pueden ser transmitidas. Estas herramientas han ofrecido el motivo para publicar en los nuevos medios, tales como los libros electrónicos, las revistas, las bibliotecas y los archivos. Junto con el desarrollo de tales recursos, una amplia variedad de herramientas de búsqueda, de recuperación y el empleo de técnicas como el hipertexto, los agentes inteligentes, los knowbots, etc. que permiten a los usuarios que busquen eficientemente la información necesaria. De forma similar, la convergencia de varios medios en una sola plataforma ha originado sistemas de información como multimedia, hipermedia y polimedia. Ahora, la pregunta es ¿ qué se puede medir en este contexto? Si nos referimos a los elementos que se mencionaron al hablar del Ciberespacio, podríamos clarificar esta circunstancia. Se proponen a continuación algunos de estos elementos:

1. El número, el alcance y los temas de las rede s de información

2. Distribución de las redes por países

3. Volumen de las colecciones de información en las redes por tamaño y tipo

4. Distribución de los diversos tipos de redes

5. Evaluación de los tiempos de respuesta de las redes y provisiones de acceso

Internet, como enorme autopista de la información, ha proporcionado argumentosmuy interesantes para el estudio. Por ejemplo para el estudio del E-mail podemos hacer los siguiente:

1. El número de direcciones de correo

2. Distribución de las direcciones de correo por países, organismos e instituciones

3. Uso del correo en los sectores público y privado

4. El volumen, el tipo y el tamaño de la información enviada a través del correo

5. Distribución de los usuarios de correo por profesiones y empresas

6. Proporciones de diversos tipos de documento enviados por correo

Éstas son las áreas, que se pueden cuantificar usando medidas estadísticas y técnicas informétricas..

Uno de los medios de información que más profundamente han influido el mundo de la información en el mundo entero es el World Wide Web, un Web de información hipertexto multimedia que opera como una de las autopistas de Internet. Esta tecnología, siempre en expansión, ha provocado cambios tanto a nivel individual como en las diversas actividades sociales. Hoy en día, todas las organizaciones, instituciones tanto públicas como privadas tienen sus propias sedes y homepages. Podemos encontrar fácilmente todas las clases y formatos de la información en el World Wide Web.

Una gran cantidad de productores, de proveedores y de vendedores de la información han puesto sus colecciones en el Web

Por lo tanto, la métrica y la medida de estos medios impresionantes, sin ninguna duda, son un área interesante para la investigación. Algunas de estas áreas de estudio son las siguientes:

1. El número de sedes Web y de homepages en el mundo y también su distribución países

2. Clasificación de las páginas Web por tipos de documentos

3. Número de páginas Web por dominios

4. Clasificación de páginas Web por el idioma de los documentos y por los modos de representación de la información

5. Estadísticas de uso y usuarios de las paginas Web en un período del tiempo dado

6. El número de citas recibidas por cada página Web

7. Ordenar los Web más citados y páginas personales según el tipo de documento

8. Los tipos de colecciones electrónicas disponibles en cada sede Web

9. Factor de Impacto del Web y productividad de los autores

10. Análisis del contenido de las páginas Web

11. Identificar la variedad de publicaciones electrónicas por el tipo, el idioma y la distribución geográfica

Estas medidas cuantitativas del Web no pueden solamente mostrar la anchura y la amplitud del WWW sino también pueden mostrar las etapas de desarrollo de los recursos del WWW a través del mundo.

Midiendo recursos electrónicos tales como libros electrónicos, revistas, bibliotecas y fuentes de referencia se pueden elaborar otras investigaciones que nos permitan reconocer la transición revolucionaria de lo impreso al mundo electrónico. Para tener una idea del análisis cuantitativo de estos recursos electrónicos, algunos de los principales aspectos que se pueden tratar son:

1. Estadística de bibliotecas digitales

2. Número de revistas electrónicas por temas e idiomas

3. Número de revistas publicadas en ambos formatos (electrónico y papel)

4. Número de fuentes de referencia electrónica disponibles

5. Análisis de citas de revistas electrónicas

6. Utilización de las revistas electrónicas

7. Distribución de recursos electrónicos por tipo, país e institución

8. Productividad científica en el entorno electrónico

9. Crecimiento de la literatura electrónica y su obsolescencia

Para llevar a cabo estas investigaciones es preciso trabajar con agentes inteligentes, robots del conocimiento así como con motores de búsqueda del Web que son herramientas eficaces para extraer apropiadamente la información relevante.

1.5. La investigación realizada.

Nuestro trabajo de investigación cibermétrico, comenzó en el año 1997 con la creación de un robot de recogida de datos (Alonso, 1997) de elaboración propia que nos permitía la recogida de los datos relativos a diferentes sedes Web para su posterior análisis y estudio. La creación de este robot se debió esencialmente a que los trabajos anteriores de tipo cibermétrico basaban sus estudios en los rastreadores de información, pero diferentes estudios (Lawrence, 1998) atestiguaban que los rastreadores solamente encontraban una pequeña fracción de todas las sedes disponibles y las variaciones en la información que cubren, en función de la accesibilidad de la información varía entre los diferentes rastreadores (Lawrence, 1999). Estos problemas han sido constatados con posterioridad por (Bergman, 2000) indicando que cada vez con mayor frecuencia los buscadores analizan menor cantidad de información, dando lugar al denominado Web invisible, que es la parte del Web no analizada. Este término ha sido acuñado por Jill Ellsworth en 1994.

El concepto de sede, que algunos investigadores (Aguillo, 2000) consideran como una nueva unidad documental adecuada para este tipo de estudios lo definen "como el conjunto de páginas Web ligadas jerárquicamente a una página principal, representables por la URL de ésta y que forman una unidad documental, distinguible de otras, y una unidad institucional, en la que es posible identificar la responsabilidad de su autoría." (Aguillo, 1998).

Nos propusimos analizar los diferentes dominios Web españoles asociados a instituciones investigadoras, principalmente sedes Web de universidades. La idea inicial era recoger los datos del mayor número posible de sedes con el fin de disponer de una gran variedad de dominios, que nos permitieran poder obtener datos mucho más fiables.

En este proceso de toma de decisión de los dominios españoles a ser analizados, surgieron problemas con el Centro de Proceso de Datos de la

Universidad de Salamanca y nos vimos obligados a solicitar permiso expreso de cada una de las sedes para poder ejecutar la recogida de datos correspondiente y ello fue un aspecto que limitó de forma extraordinaria las pretensiones iniciales de la investigación. Solamente fue posible realizar la recogida de datos de aquellos dominios españoles que dieron este permiso de forma expresa. Por ello la selección de dominios puede verse en algún caso sesgada, desde el punto de vista geográfico.

Finalmente los dominios Web españoles que se pudieron rastrear y sobre los que se pudieron recoger datos fueron los siguientes:

- cervantes: Instituto Cervantes.

- cicyt: Centro de Investigación en Ciencia y Tecnología.

- ciemat: Centro de Investigaciones Energéticas, Medioambientales y tecnológicas.

- deusto: Universidad de Deusto.

- esa: Agencia Espacial Europea.

- fundesco: Fundación para el Desarrollo de la Función Social de las Comunicaciones.

- hrc: Hospital Ramón y Cajal.

- impi: Instituto de la Pequeña y Mediana Empresa.

- uam: Universidad Autónoma de Madrid.

- uc3m: Universidad Carlos III de Madrid.

- ugr: Universidad de Granada.

- ujaen: Universidad de jaén.

- uji: Universidad Jaume I de Castellón.

- um: Universidad de Murcia.

- unex: Universidad de Extremadura.

- unican: Univesidad de Cantabria.

- unnet: Universidad Antonio de Nebrija.

- upco: Universidad Pontificia de Comillas.

- upna: Universidad Pública de Navarra.

- upsa: Universidad Pontificia de Salamanca.

- url: Universidad Ramón Llul.

- urv: Universidad Rovira i Virgili.

- us: Universidad de Sevilla.

- usal: Universidad de Salamanca.

- uv: Universidad de Valencia.

- uva: Universidad de Valladolid.

- vhebron: Hospital Vall d'Hebron.

Se realizaron tres recogidas de datos en Julio de 1998, Junio de 1999 y Enero de 2000.

Para el estudio de todos los dominios se ha optado por hacerlo de la siguiente manera:

1. Análisis cuantitativo de los datos, según se recoge en el capítulo 3.

2. Análisis de diferentes medidas topológicas, que nos permitirán obtener diferentes valores de similaridad, que nos ofrecerán unos buenos indicadores de la variación entre las diferentes recogidas. Se analiza en el capítulo 4.

3. Análisis de las leyes de exponenciación, nueva corriente de estudio, en la que nos adentramos e intentamos ofrecer las características españolas.

Para el trabajo con los datos obtenidos y almacenados en una base de datos relacional se han utilizado diferentes planteamientos, que pasamos a enumerar:

1) La mayor parte del tratamiento de los datos se ha realizado mediante la creación de sentencias SQL que nos han devuelto los resultados en tablas que hemos podido manejar de forma directa. En estos casos se ha trabajado con la totalidad de los datos recogidos.

2) Cuando hemos necesitado trabajar con procesamiento de matrices, hemos trabajado con los mil primeros nodos de cada uno de los dominios y dentro del trabajo, cuando se trabaja en estas condiciones se indica expresamente.

En este sentido conviene realizar una aclaración del porqué de esta decisión. La mayor parte de los trabajos existentes, que utilizan las matrices para realizar sus cálculos, partiendo de recogidas de datos con millones de documentos, luego solamente utilizan unos pocos cientos de los mismos.

(Ingwersen, 1998) de algunos millones de documentos se queda solamente con 200, recogidos de forma aleatoria. En nuestro caso el centrarnos en mil documentos viene determinado por intentar trabajar con el mayor número de documentos posibles, pero atendiendo a las posibilidades de procesamiento que estaban a nuestro alcance. Para ello realizamos una simulación del tiempo de procesamiento que precisábamos para las matrices, teniendo en cuenta que el número de matrices era alto para cada dominio y para cada recogida. En esta simulación el punto adecuado rondaba los mil documentos.

Con los datos de estos mil documentos generamos la matriz en formato de importación Matlab y se crearon diferentes programas para el tratamiento de dichas matrices. El equipo que se empleó para el procesamiento de los datos fue el siguiente:

Hardware:

SUNW Ultra SPARC-II (296 Mhz)

RAM: 128 Mb

Software:

SunOS Release 5.5.1. Version Generic [UNIX(R) System V Release 4.0]

3) Para el procesamiento de las leyes de exponenciación los datos se obtuvieron con Matlab y se pasaron posteriormente al programa StatGraphics, procesando y obteniendo los gráficos de todas las leyes para cada uno de los dominios y de las recogidas, conformando 324 gráficos que representan las leyes. De todos estos gráficos vamos a incluir en nuestro trabajo solamente el que mejor y peor resultado ofrecen en cada ley y en cada recogida.

2. Análisis Cuantitativo.

Como ya vimos en el capítulo 1, por las gráficas del desarrollo tanto de Internet como del Web, el crecimiento del mismo es claramente exponencial, con un éxito innegable que lo convierten en un sistema ideal para el estudio y la aplicación de las técnicas cibermétricas.

Autores como (Abraham, 1997), (Pirolli, 1996), (Pitkow, 1998) y (Adamic, 1999) constataron la necesidad de aplicar nuevas medidas e interpretaciones en los intentos de medir e interpretar la estructura, tamaño y conectividad del Web, en constante evolución y con una alta volatilidad (Koehler, 1999), (Koehler, 1999b).

Los primeros trabajos cuantitativos relacionados con el Web trataron fundamentalmente de estudiar la evolución del tamaño y la descripción de los primeros motores de búsqueda, tratando de conocer la entidad y cobertura de dichos motores. De esta época son los trabajos de (Mauldin, 1994), (Mauldin, 1995), (Ciolek, 1997), (Clarke, 1997) y (Sonnenreich, 1998). Más recientemente (Bar, 2000) emplea los buscadores en una investigación que trataba de comprobar si el Web podía resultar una buena fuente de datos para la investigación.

Otros estudios utilizaron los datos recogidos por algunos motores para realizar un estudio de estas características, como fueron los trabajos de (Woodruff, 1996), basado en la recogida de datos del motor Inktomi, y de (Bray, 1996), con el motor Open Text. Este último autor realizaba una buena clasificación de los diferentes aspectos que se deberían tratar, además de indicar la importancia de las sedes e incluía alguna de la terminología que posteriormente otros autores utilizan o redefinen y mejoran.

Sin embargo, en estos primero trabajos no se tenían en cuenta ni la cobertura global del Web ni su naturaleza hipertextual, que evidentemente han modificado los planteamientos en los trabajos posteriores.

Uno de los primeros trabajos que tiene en cuenta la naturaleza hipertextual del Web y que permitió la aplicación de técnicas bibliométricas fue el trabajo de (Larson, 1996). Los trabajos de (Harter, 1996) y (Harter, 1996b) sobre el estudio de los enlaces establecidos entre revistas convencionales y las revistas electrónicas supuso un nuevo planteamiento en la aplicación de estas técnicas. El concepto de *sitation*, introducido por (McKiernan, 1996) y adoptado por (Aguillo, 1996) y por

(Rousseau, 1997) introdujo un nuevo elemento previo al desarrollo de diferentes indicadores cibermétricos.

Comenzaremos nuestro estudio analizando la evolución de determinados elementos, como son algunos tipos de ficheros (compresión, gráficos, vídeo, sonido, texto, estilos, vrml, formatos Web), para pasar posteriormente a analizar la evolución de algunos tipos de etiquetas, tipos de servidores, exclusión de robots.

Posteriormente veremos el tamaño del Web, analizando diferentes aspectos y finalmente nos centraremos en la estructura hipertextual del Web, analizando las características de los enlaces.

2.1. Evolución en los tipos de ficheros.

Un aspecto a estudiar es la evolución que se produce en el Web en el empleo de determinados tipos de ficheros. Nos van a dar una indicación de las tendencias que se producen en un determinado momento en el empleo de un tipo de formato concreto, que en ocasiones se trata de tendencias puntuales y en otros casos nos dan una indicación de la implantación y del éxito de una determinada tecnología.

Se estudia para ello la evolución de los siguientes tipos:

2.1.1. Ficheros de compresión.

El número de formatos de compresión que se pueden analizar es elevado, como indica (Woodruf, 1996), aunque nosotros nos centraremos en los formatos TAR, Z y ZIP. Los datos de (Woodruff, 1996) no nos sirven como comparación, al centrarse en los datos de un único año y en una recogida puntual, además de no recoger los datos del formato TAR.

La evolución ha sido hacia un incremento en el empleo del formato ZIP, con un incremento espectacular desde el 0,1% de la primera recogida hasta un 76,9% de la segunda, para finalmente retroceder de forma moderada en la tercera recogida en beneficio del formato Z.

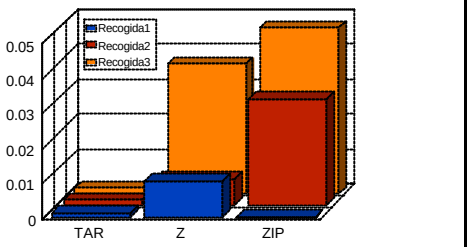


Figura 8: Formatos de Compresión en cada recogida.

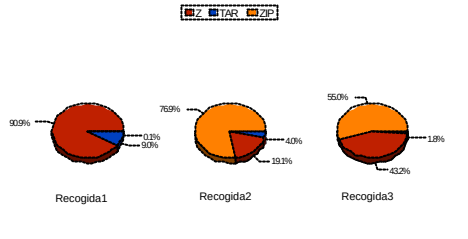


Figura 9: Evolución de los formatos de Compresión. Los porcentajes se calculan sobre el total de ficheros comprimidos.

2.1.2. Ficheros Gráficos.

En los formatos gráficos nos hemos centrado en los que se pueden utilizar en el Web, a pesar de existir un número realmente muy elevado de formatos de tipo gráfico.

De los formatos gráficos analizados claramente el formato GIF es el mayoritariamente empleado, con un ligero retroceso en la segunda recogida hacia el formato JPEG, y con un nuevo repunte en la tercera recogida, que lo convierte en el formato gráfico más empleado. En el estudio de (Woodruff, 1996) el formato GIF dominaba claramente los formatos gráficos con un 62% quedando para otros formatos unos restos mínimos y poco significativos.

En el momento de comenzar esta investigación el formato PNG no estaba todavía implantado, aunque en la última recogida sí se empleaba. Pero se decidió mantener en todas las recogidas las mismas características para que los datos fueran consistentes.

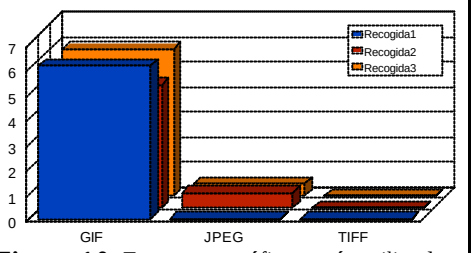


Figura 10: Formatos gráficos más utilizados.

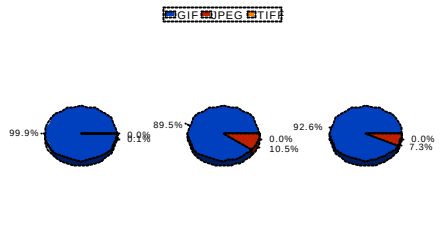


Figura 11: Evolución de los formatos gráficos. Los porcentajes se calculan sobre el total de ficheros gráficos.

2.1.3. Ficheros de Vídeo.

En los formatos de vídeo analizados las variaciones existentes en las tres recogidas son muy altas, cambiando de una recogida a otra el formato que mayoritariamente se implanta.

En la primera recogida tenemos que los formatos AVI y MPEG son los mayoritariamente empleados, para sufrir un profundo cambio en la segunda recogida, en la que el formato AVI desciende mucho y el formato MPEG aumenta considerablemente y dando entrada al formato MOV. En la tercera recogida la situación se estabiliza para los formatos AVI y MOV, que prácticamente tienen la misma cuota de utilización.

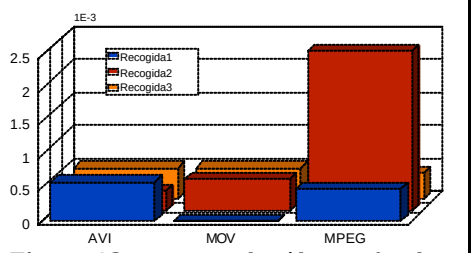


Figura 12: Formatos de vídeo analizados.

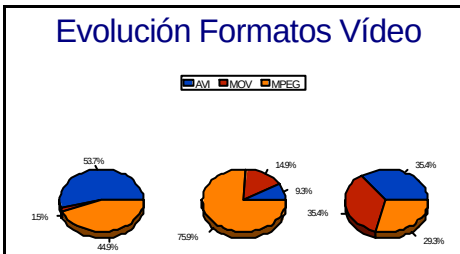


Figura 13: Evolución de los formatos de vídeo.Los porcentajes se calculan sobre el total de ficheros de vídeo.

2.1.4. Ficheros de Sonido.

En el caso de los formatos de sonido, el formato que mayoritariamente se empleaba en la primera recogida era el formato WAV, con un 92% de cuota, para descender vertiginosamente en las otras dos recogidas, en las que aumentan el resto de formatos y sobre todo el formato AU.

El formato MIDI mantiene aproximadamente su cuota entre la segunda y la tercera recogida y prácticamente no se empleaba en la primera recogida.

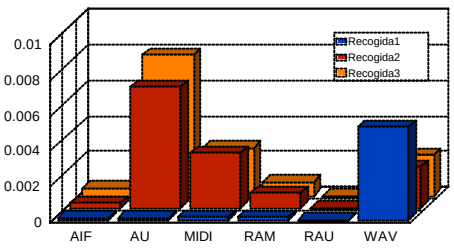


Figura 14: Formatos de sonido más utilizados.

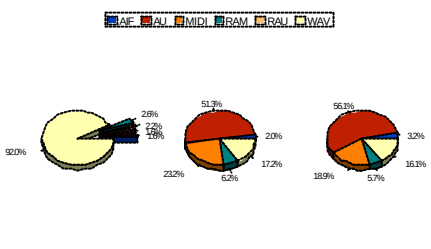


Figura 15: Evolución de los formatos de Sonido.Los porcentajes se calculan sobre el total de ficheros de sonido.

2.1.5. Ficheros de Texto.

Hay una clarísima tendencia a utilizar el formato PDF, que sufre un notable

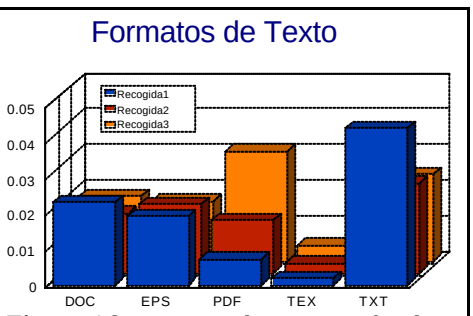


Figura 16: Formatos de Texto analizados.

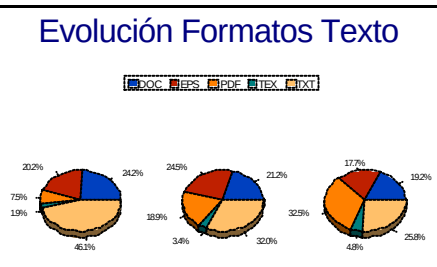


Figura 17: Evolución de los formatos de Texto.Los porcentajes se calculan sobre el total de ficheros de texto.

incremento desde la primera recogida, aumentando también su cuota, aunque de forma mucho más moderada el formato TEX.

Los formatos DOC y TXT sufren continuos retrocesos en su utilización.

2.1.6. Utilización de Estilos.

Los ficheros de estilo permiten crear una definición de diferentes estilos que se pueden aplicar al texto de las páginas Web, identificando cada estilo mediante una palabra única. Para aplicar dicho estilo predefinido de antemano, basta con aplicar el nombre asignado a dicho estilo.

Los estilos pueden definirse dentro de la propia página Web o en un fichero aparte al que se hace referencia mediante un link. Nosotros recogemos los datos de las dos modalidades y por lo tanto los datos obtenidos hacer referencia al global de la utilización de estilos.

El empleo de ficheros de estilo ha sufrido una notable evolución desde la primera recogida hasta la última, con un incremento muy importante.

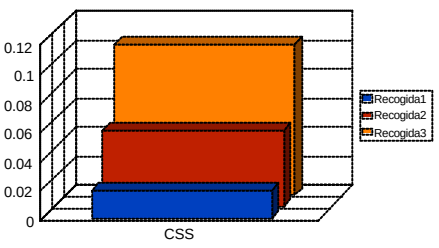


Figura 18: Empleo de Estilos.

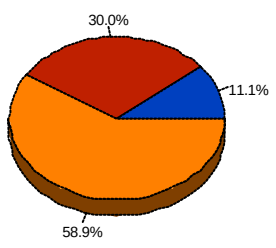


Figura 19: Evolución de los estilos en las tres recogidas.Los porcentajes se calculan sobre el total de las recogidas.

2.1.7. Utilización de VRML.

El **VRML** (Virtual Reality Modeling Language: lenguaje de modelación de la realidad virtual) es un lenguaje de programación con el que se pueden desarrollar mundos interactivos en tres dimensiones (3-D). Estos mundos constituyen lo que

se denomina la "realidad virtual", porque los usuarios pueden interactuar con los objetos de una forma similar a como lo hacen en la realidad "normal".

La **realidad virtual** puede revolucionar la manera con la que los usuarios se relacionan con sus ordenadores, de un modo similar a lo ocurrido con el World Wide Web. Las posibilidades son innumerables: simulaciones educativas, nuevos métodos de organizar la información, nuevas formas de entretenimiento, etc.

El empleo de ficheros VRML aumenta entre la primera y la segunda recogida, para descender su empleo en la tercera.

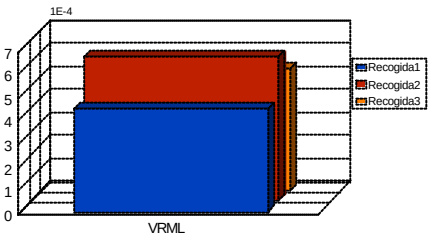


Figura 20: Utilización del formato VRML.

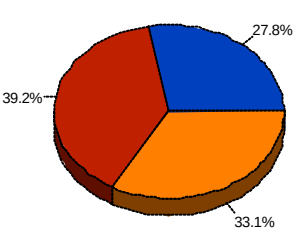


Figura 21: Evolución del empleo de VRML entre las tres recogidas. Los porcentajes son sobre el total de ficheros VRML en las tres recogidas.

El aumento coincidió con lo que parecía ser la implantación de un nuevo estándar, pero que no llegó a cuajar como sistema para poder representar animaciones.

2.1.8. Ficheros Web.

El empleo de formatos para el Web se decanta clarísimamente por el formato HTML frente a ASP, que aumenta muy poco entre las tres recogidas.

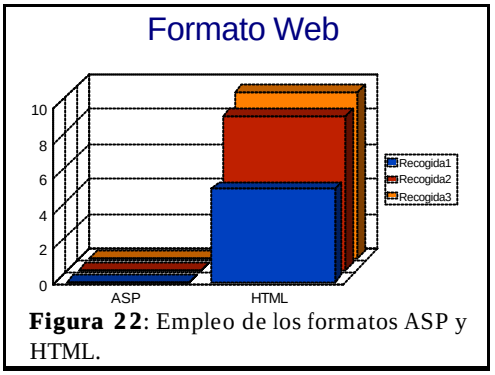


Figura 22: Empleo de los formatos ASP y HTML.

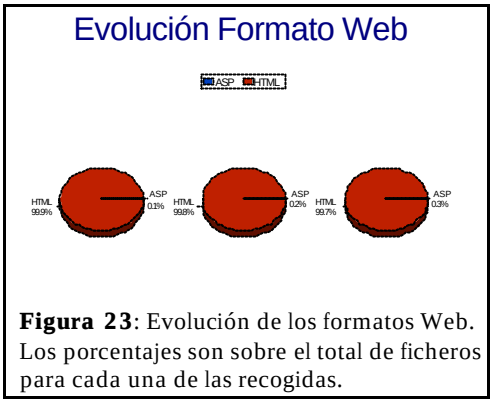


Figura 23: Evolución de los formatos Web. Los porcentajes son sobre el total de ficheros para cada una de las recogidas.

2.1.9. Evolución Multimedia.

Para finalizar con este apartado, valoramos la evolución multimedia existente en los dominios analizados. Para ello hemos tenido en cuenta, los datos correspondientes a los formatos gráficos, de vídeo, de sonido y VRML. Como podemos ver en el gráfico no existe una utilización desmesurada de elementos multimedia y la evolución en este sentido es mínima.

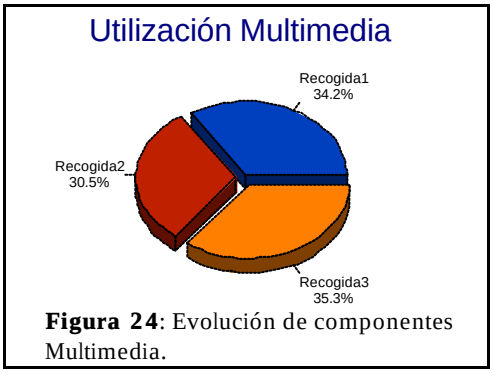


Figura 24: Evolución de componentes Multimedia.

2.2. Empleo de etiquetas.

Vamos a analizar la utilización de determinas etiquetas, su evolución, y en algunos casos las opciones más utilizadas de algunas de ellas.

2.2.1. Etiqueta Title.

Se emplea mayoritariamente esta etiqueta, que en las dos últimas recogidas supera el 90% de utilización, aunque no podemos conocer si el empleo de dicha etiqueta es el adecuado o no..

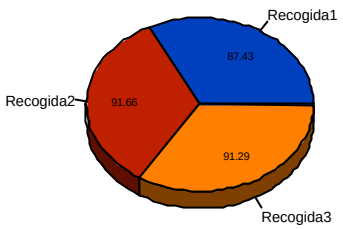


Figura 25: Porcentaje de uso de la etiqueta de Título.

2.2.2. Etiquetas varias.

En la imagen siguiente podemos ver la evolución de algunas etiquetas, y los datos más interesantes serán comentados a continuación.

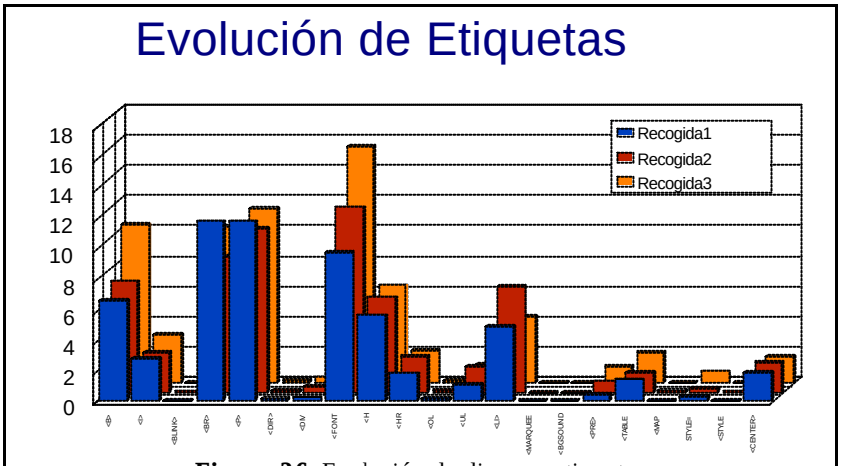


Figura 26: Evolución de diversas etiquetas.

- Uno de los datos que podemos observar es que se ha cambiado el uso de la etiqueta BR por el de la etiqueta P, que se emplea más. Esto se acentúa con el empleo de editores HTML, que suelen poner la etiqueta P de forma predeterminada.
- Se emplea mucho más la etiqueta FONT que la etiqueta H, sufriendo la etiqueta FONT un aumento considerable con el paso del tiempo.
- Se emplean más las listas sin ordenar que las ordenadas, aunque se encuentran en desuso.
- La etiqueta TABLE se emplea cada vez más y sobre todo el aumento se produce en la tercera recogida.
- Las etiquetas de estilo se emplean también cada vez más, aumentando mucho en la tercera recogida que confirma los datos obtenidos en el apartado 5.2.4.6.
- Significar la escasa incidencia de la etiqueta MAP

2.2.3. Opciones de la etiqueta Body.

En la etiqueta BODY se utiliza mayoritariamente la opción BGCOLOR frente a BACKGROUND, aumentando esta diferencia en la segunda recogida.

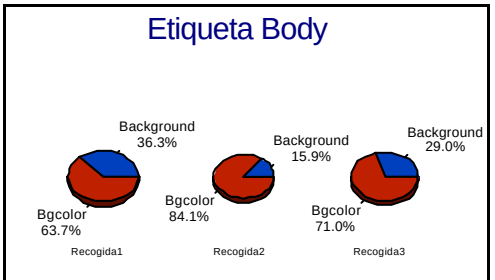


Figura 27: Porcentaje de uso de las opciones de BODY.

2.2.4. Utilización de Applet y Script.

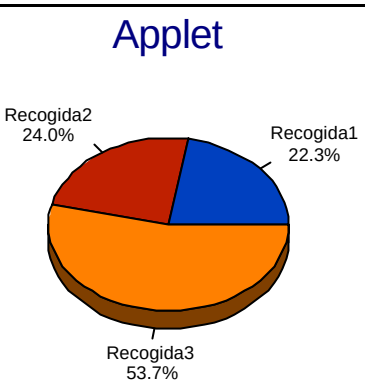


Figura 28: Porcentaje de uso de Applets entre recogidas.

Los applets son pequeñas aplicaciones escritas en lenguaje Java, y desarrolladas para ejecutarse específicamente desde un visor de HTML. Estas aplicaciones aúnan las ventajas de la tecnología Java (portabilidad del código, relativa sencillez del lenguaje, posibilidades multimedia, etc.) con las derivadas de la gran difusión de los visores Web.

Hay un claro aumento en la utilización de los applets, sobre todo en la tercera recogida con un aumento muy importante.

Parece lógico el aumento de los mismos ya que muchos de los programas editores añaden el código de forma automática y transparente para el usuario permitiendo añadir efectos de gran vistosidad sin que el usuario tenga la necesidad de conocer el lenguaje de programación. También debido a la potencia que pueden ofrecer a las páginas Web, convirtiéndose en un elemento cada vez más habitual en las páginas desarrolladas.

Los scripts se basan en el empleo de lenguaje JavaScript, que sintácticamente es muy similar a C++ o a Java, aunque presenta importantes diferencias con estos lenguajes derivadas de su naturaleza interpretada y en las que no entraremos. Inicialmente el desarrollo de los scripts lo planteó la compañía Netscape, dándole el nombre de JavaScript, aunque posteriormente Microsoft realizó una modificación del mismo llamada Jscript y finalmente desarrolló su propio lenguaje llamado VBScript.

En lo referente al empleo de scripts, su utilización es muy pareja en las tres recogidas de datos, teniendo valores muy similares.

En este caso las variaciones se producen en el lenguaje utilizado en el script, con un descenso progresivo en el empleo de java, a favor de otros lenguajes.

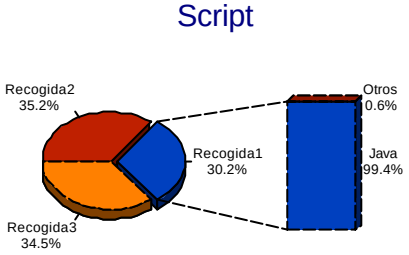


Figura 29: Empleo de Lenguajes en la 1ª Recogida.

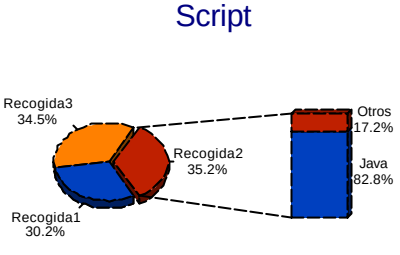


Figura 30: Empleo de Lenguajes en la 2ª Recogida.

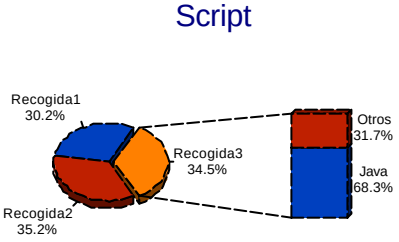


Figura 31: Empleo de Lenguajes en la 3ª Recogida.

2.2.5. Utilización de Frames.

El empleo de frames o marcos (varias ventanas independientes formando un conjunto) ha descendido considerablemente desde la primera recogida, y recuperándose levemente en la última.

Este descenso puede deberse a la complejidad añadida del manejo de

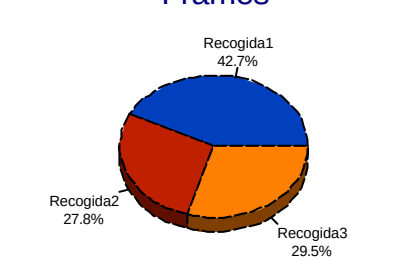


Figura 32: Evolución en el empleo de Frames.

diferentes ventanas en un mismo entorno, que complica ligeramente la realización de las páginas Web.

El empleo de frames se ha sustituido de forma habitual por el de tablas, más cómodas de utilizar y que se han convertido en un buen sistema para realizar la maquetación de una página Web y además una vez realizad una plantilla de la misma es fácilmente reutilizable y con un coste menor en el desarrollo de las páginas Web.

2.2.6. Utilización de Formularios.

Los formularios son un elemento que ha comenzado a utilizarse con bastante frecuencia en las páginas Web y vamos a ver su evolución entre las tres recogidas, así como algunas de las posibilidades de empleo de dicho elemento diferenciando cunado se emplean con el método POST o con el método GET y cuando se han utilizado como mecanismo de envío a cuentas de correo directamente.

Respecto a la utilización de formularios ha aumentado en todas las recogidas de forma importante, indicando que es un sistema bastante asumido por los usuarios.

De su utilización podemos destacar, que se utiliza menos para enviar formularios directamente a una cuenta de correo, con un descenso importante en la segunda recogida y que el método POST es el mayoritariamente empleado, muy destacado en la tercera recogida.

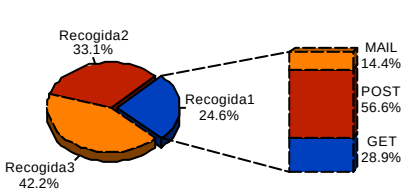


Figura 33: Empleo de formularios en la 1ª Recogida.

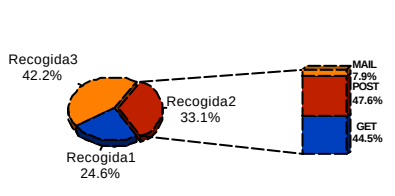


Figura 34: Empleo de formularios en la 2ª Recogida.

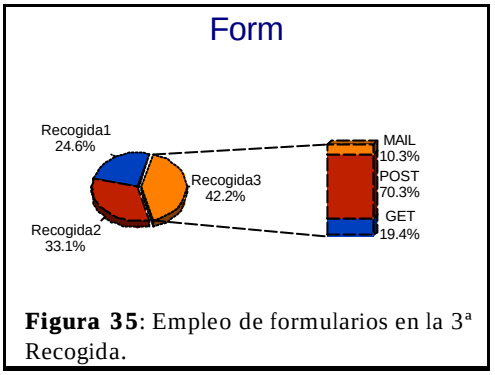


Figura 35: Empleo de formularios en la 3ª Recogida.

2.3. Utilización de Servidores Web.

El empleo de servidores Web tiene una clara tendencia a emplear el servidor Apache, que aumenta su uso en cada una de las recogidas de forma clara, con más de un 50% de máquinas con este servidor instalado.

Servidores como CERN o NCSA que en la primera recogida tenían una buena aceptación han dejado su cuota y han perdido posiciones de forma clara.

Los servidores de Microsoft han aumentado su cuota de empleo, pero bastante alejados del servidor Apache. Los servidores Microsoft mantienen una baja cuota de empleo, con ligeras modificaciones entre las recogidas.

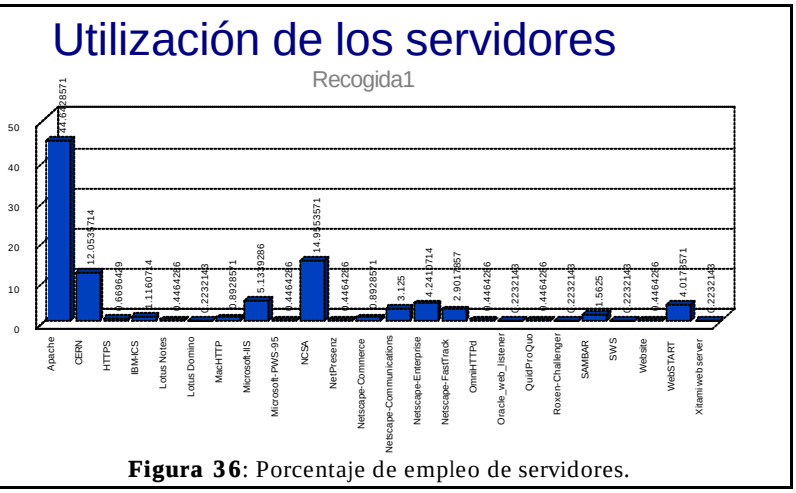


Figura 36: Porcentaje de empleo de servidores.

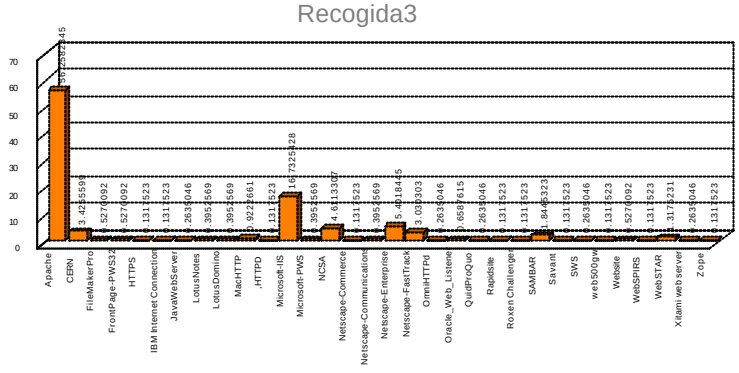


Figura 37: Porcentaje de empleo de servidores.

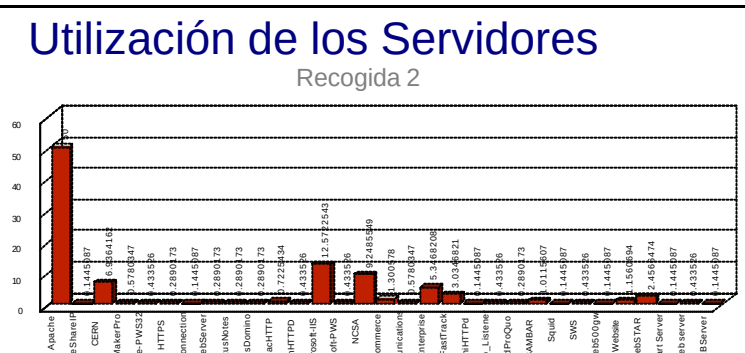


Figura 38: Porcentaje de empleo de servidores.

2.4. Exclusión de Robots.

Uno de los aspectos que no se ha estudiado con anterioridad y que nosotros vamos a valorar en nuestro trabajo, es la existencia del denominado Estándar de Exclusión de Robots (en adelante SRE) (Koster, 1994), generado el 30 de Junio de 1994 en la lista sobre robots *robots-request@nexor.co.uk* (Koster, 1994). Todo los relacionado con robots puede analizarse en (Koster, 1993), (Koster, 1993b), (Koster, 1994), (Koster, 1995), (Koster, 1996), (Koster, 1996b).

En este momento el SRE se basa en dos implementaciones, complementarias entre sí:

- Protocolo de exclusión de robots.
- Etiqueta META para robots.

2.4.1. Protocolo de exclusión de robots

Este protocolo (Koster, 1994) se basa en la existencia de un fichero *robots.txt* en el raíz del servidor Web. En dicho fichero se incluyen las directivas que indican al robot las operaciones a realizar en el servidor Web con el que está trabajando.

Básicamente, el contenido del fichero *robots.txt* es el siguiente:

- Una línea User-agent: en la que se indica el nombre del robot que se verá afectado por las directivas que vayan a continuación. Si se pone un * afectará a todos los robots.
- Un número indeterminado de líneas con la directiva Disallow: en la que se indica el path que se desea restringir. Es importante indicar que el protocolo exige una línea por cada path restringido. Si esta directiva se encuentra en blanco, indicará que no se aplica ninguna restricción.

Un ejemplo de fichero sería el siguiente:

User-agent: *

Disallow:

User-agent: Sonda Ciberdocumental

Disallow: /bibesp

Disallow: /logs

User-agent: lycos

Disallow: /

2.4.2. Etiqueta META para robots

Esta implementación es muy posterior al protocolo (existen unas notas preliminares de mayo de 1996 en la dirección Web:)

(http://info.webcrawler.com/mak/projects/robots/meta-notes.html)

y se basa en la indicación de permisos en cada página HTML concreta, permitiendo a cada desarrollador de páginas Web indicar sus propias indicaciones, independientes de la política que pueda tener el administrador de un determinado servidor.

Su utilización se basa en el empleo de la etiqueta META en la sección HEAD del documento HTML.

Las opciones que afectan son:

- Opción *name*, que contendrá la palabra robots
- Opción *content*, que puede contener las directivas all, none, index, noindex, follow y nofollow.

Unos ejemplos serían:

< meta name = ”robots” content = ”index, follow“ >

< meta name = ”robots” content = ”index, nofollow“ >

< meta name = ”robots” content = ”none“ >

2.4.3. Situación del SRE en los dominios analizados.

2.4.3.1. Protocolo de Exclusión de Robots

La evolución en el empleo de dicho protocolo, ha sido creciente desde la primera a la última recogida de datos, si bien, la implantación del protocolo de exclusión de robots es escasa, como podemos ver en la Figura 39 .

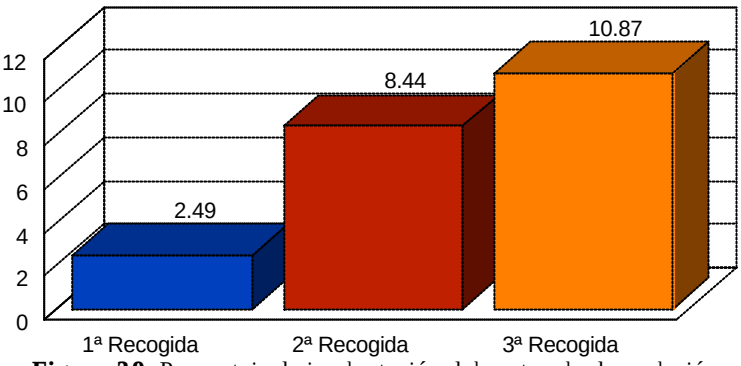


Figura 39: Porcentaje de implantación del protocolo de exclusión.

En cuanto, al modo de utilización del protocolo, podemos indicar que en la primera recogida de datos, los servidores que emplean bien el protocolo son el 89% y de éstos el 100% es aplicable a todos los robots.

En la segunda recogida de datos, el número de dominios que emplean el protocolo se duplica (14), los servidores que emplean bien el protocolo son el 80%. De estos el 95% afecta a todos los robots y un 11% a robots particulares.

En la tercera recogida de datos, el número de dominios que emplean el protocolo aumenta ligeramente (16), los servidores que emplean bien el protocolo son el 64%. De estos el 100% afecta a todos los robots y un 24,4% a robots particulares.

El número medio de directorios excluidos es de 2,4 (con un mínimo de 2 y un máximo de 4); 3,7 (con un mínimo de 1 y un máximo de 19) y 4,1 (con un mínimo de 1 y un máximo de 15) respectivamente.

2.4.3.2. Etiqueta META para robots

En la primera recogida no hay datos.

En la segunda recogida de datos, el 1,57% de las páginas emplea este método de exclusión, de éstas el 56,12% se emplea para excluir. De las páginas con exclusión el 83,54% utiliza la directiva nofollow y el 16,46% la directiva none.

En la tercera recogida de datos, el 1,01% de las páginas emplea este método de exclusión, de éstas el 0,89% se emplea para excluir. De las páginas con exclusión el 45,16% utiliza la directiva nofollow y el 54,84% la directiva none.

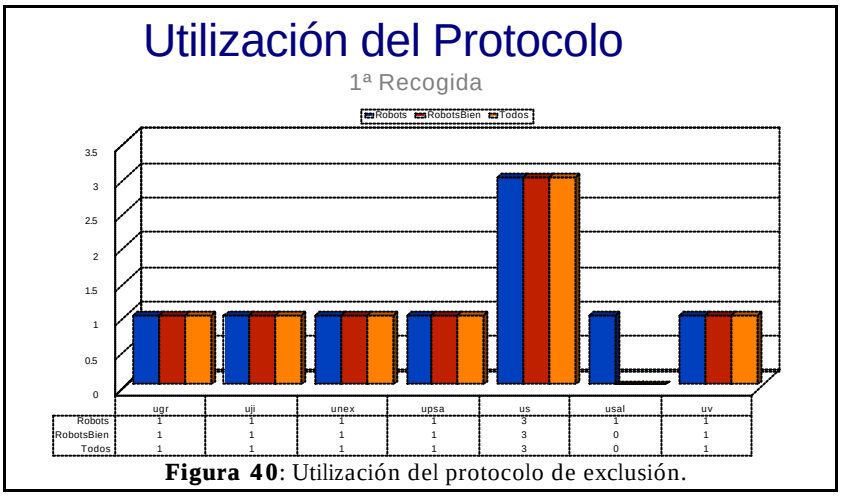


Figura 40: Utilización del protocolo de exclusión.

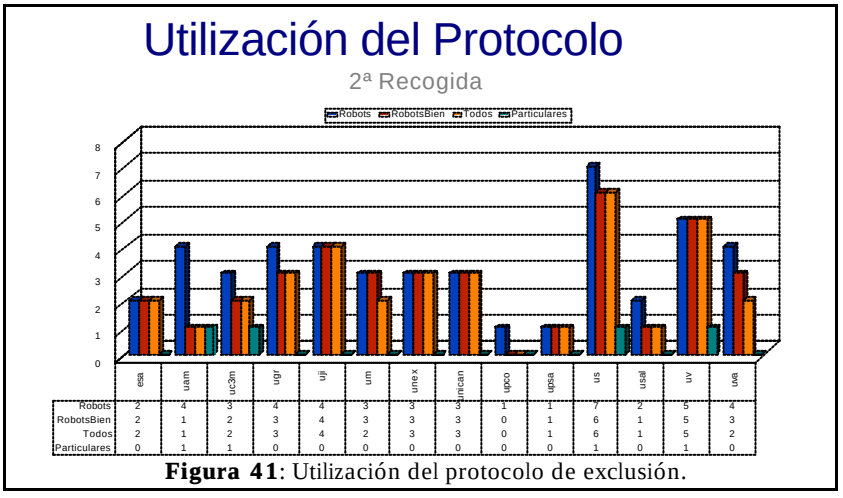


Figura 41: Utilización del protocolo de exclusión.

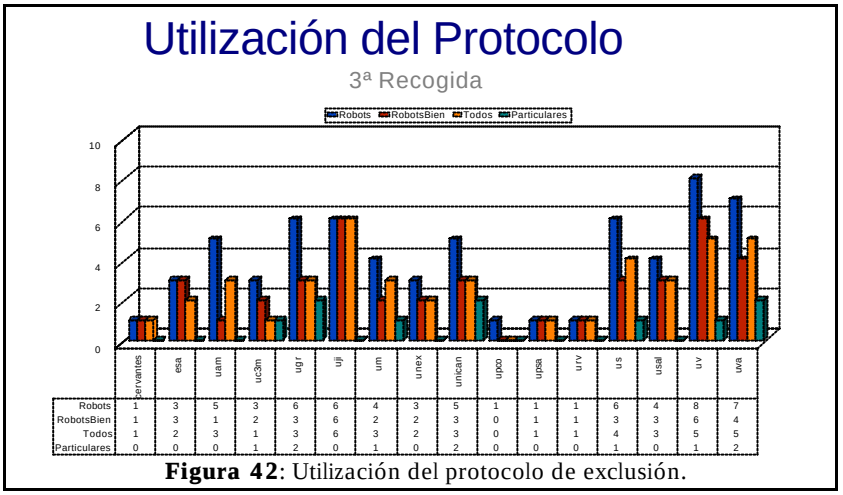


Figura 42: Utilización del protocolo de exclusión.

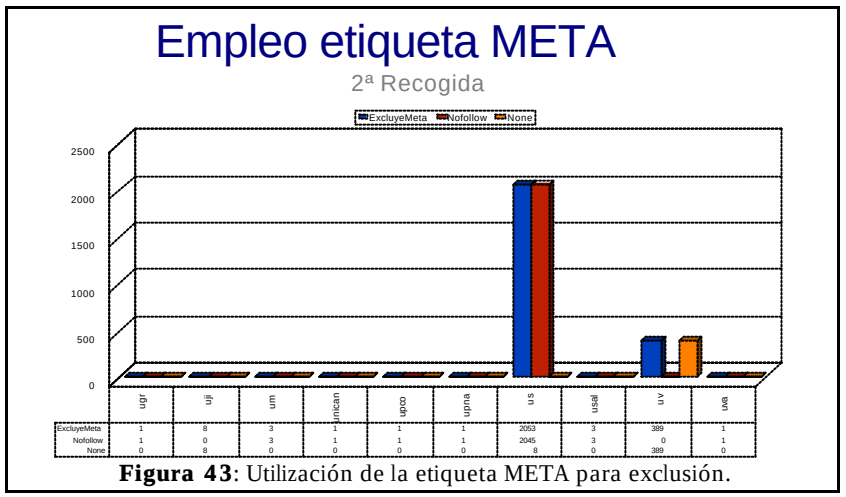


Figura 43: Utilización de la etiqueta META para exclusión.

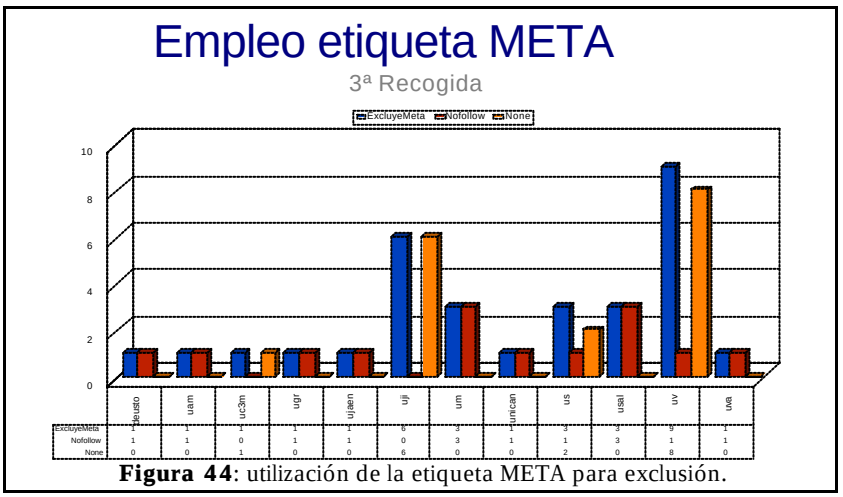


Figura 44: utilización de la etiqueta META para exclusión.

2.5. Tamaño del Web.

2.5.1. Número único de nodos (de páginas).

El número de nodos o tamaño documental (Aguillo, 2000) es considerado por algunos autores como un indicador cibermétrico de primer orden (Aguillo, 2000) y ofrece una buen estimación del crecimiento de las diferentes sedes, que nos permite valorar si los dominios tienen o no un fuerte crecimiento.

Podemos ver una importante evolución en el número de los nodos, entre la primera recogida y la segunda, pero muy leve entre la segunda y la tercera.

Entre la primera recogida y la segunda se produjo el boom del World Wide Web y se refleja perfectamente en los datos obtenidos, mientras que entre la segunda y la tercera fue un periodo de mayor tranquilidad.

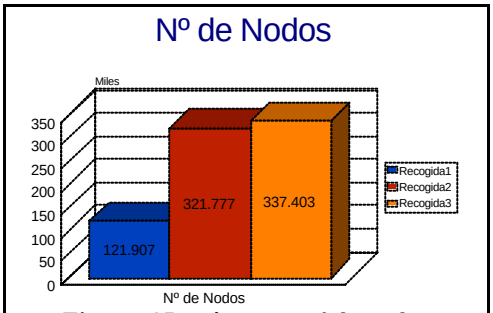


Figura 45: Número total de nodos.

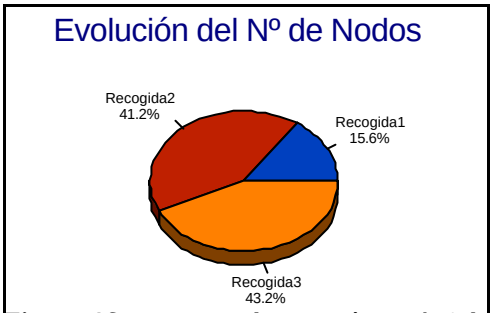


Figura 46: Porcentaje de variación en el nº de nodos.

Analizando individualmente cada uno de los dominios, el porcentaje de evolución de cada dominio se refleja en el siguiente gráfico.

Como podemos ver, la evolución individual de cada nodo es importante, sobre todo de la primera recogida a la segunda, con alguna excepción, siendo ligeramente más moderada en la tercera recogida, con algunos dominios que tienen un notable incremento en esta última recogida.

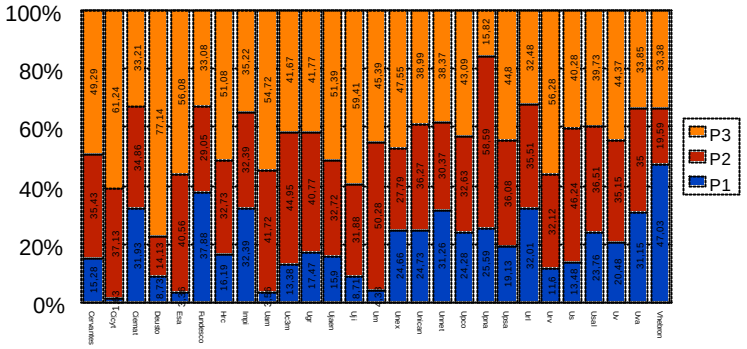


Figura 47: Porcentaje de variación nº nodos por dominios.

Diferenciando el porcentaje que ocupa cada dominio sobre el global de la recogida los datos son los siguientes:

Dominio	Porcen-R1
Uv	19,21
Uva	16,82
Unican	10,41
Urv	7,33
Us	6,94
Ugr	6,28
Usal	5,39
Uc3m	5,39
Unex	4,23
Uji	3,59
Uam	2,51
Url	1,84
Cervantes	1,65
Vhebron	1,59
Ciemat	1,31
Um	0,97
Ujaen	0,92
Fundesco	0,69
Upna	0,66
Upco	0,66

Dominio	Porcen-R2
Uv	12,49
Uam	11,11
Us	9,02
Urv	7,68
Uva	7,16
Uc3m	6,86
Unican	5,78
Ugr	5,55
Uji	4,98
Um	4,25
Usal	3,14
Unex	1,8
Cervantes	1,45
Esa	1,36
Url	0,77
Ujaen	0,72
Upna	0,58
Ciemat	0,54
Deusto	0,36
Upco	0,34

Dominio	Porcen-R3
Uv	15,03
Uam	13,9
Urv	12,84
Uji	8,84
Us	7,49
Uva	6,6
Uc3m	6,06
Unican	5,93
Ugr	5,43
Um	3,66
Usal	3,26
Unex	2,94
Cervantes	1,92
Deusto	1,89
Esa	1,8
Ujaen	1,08
Url	0,67
Cicyt	0,52
Ciemat	0,49
Upco	0,42

Deusto	0,59
Impi	0,37
Esa	0,3
Upsa	0,16
Unnet	0,14
Hrc	0,04
Cicyt	0,04

Cicyt	0,33
Vhebron	0,25
Fundesco	0,2
Impi	0,14
Upsa	0,11
Unnet	0,05
Hrc	0,03

Vhebron	0,41
Fundesco	0,22
Upna	0,15
Impi	0,14
Upsa	0,13
Unnet	0,06
Hrc	0,04

2.5.2. Número único de servidores.

En cuanto al número de servidores las tendencias son similares a lo ocurrido con el número de nodos, es decir, una mayor evolución entre la primera y la segunda recogida y algo menor entre la segunda y la tercera.

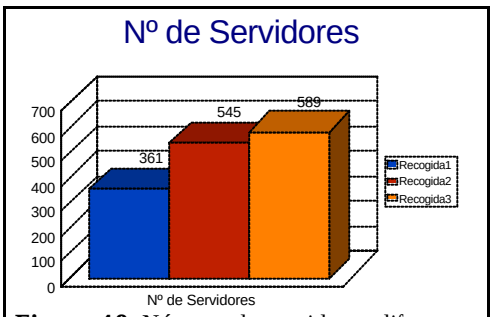


Figura 48: Número de servidores diferentes.

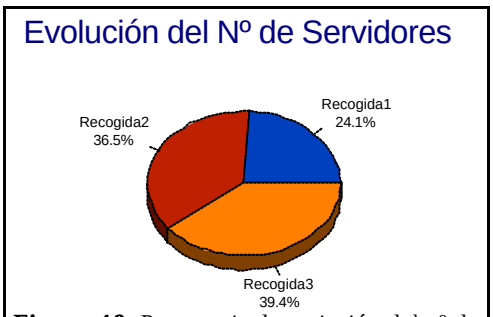


Figura 49: Porcentaje de variación del nº de servidores.

Analizando individualmente cada dominio el porcentaje de variación se refleja en el siguiente gráfico.

Podemos observar 4 dominios que no modifican su número de servidores en ninguna de las recogidas, aunque en general el número de servidores se modifica en cada recogida.

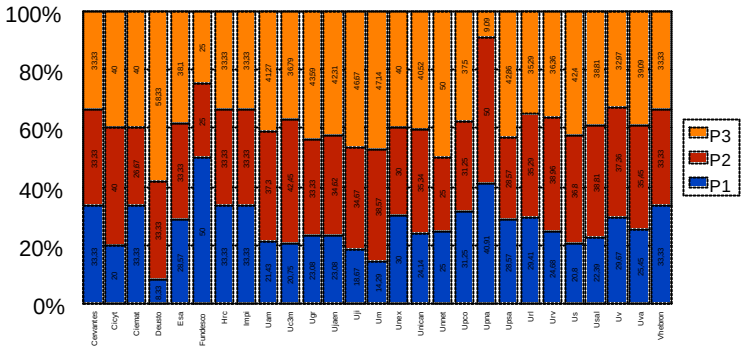


Figura 50: Porcentaje de variación nº servidores por dominio

Diferenciando el porcentaje que ocupa cada dominio sobre el global de la recogida los datos son los siguientes:

Dominio	Porcen-R1
Uva	15,51
Uv	14,96
Unican	7,76
Uam	7,48
Ugr	7,48
Us	7,2
Uc3m	6,09
Urv	5,26
Unex	4,16
Usal	4,16
Uji	3,88
Um	2,77
Upna	2,49
Ujaen	1,66
Esa	1,66
Url	1,39
Upco	1,39

Dominio	Porcen-R2
Uva	14,31
Uv	12,48
Uam	8,62
Us	8,44
Uc3m	8,26
Unican	7,52
Ugr	7,16
Urv	5,5
Um	4,95
Uji	4,77
Usal	4,77
Unex	2,75
Upna	2,02
Ujaen	1,65
Esa	1,28
Url	1,1
Upco	0,92

Dominio	Porcen-R3
Uva	14,6
Uv	10,19
Us	9
Uam	8,83
Ugr	8,66
Unican	7,98
Uc3m	6,62
Uji	5,94
Um	5,6
Urv	4,75
Usal	4,41
Unex	3,4
Ujaen	1,87
Esa	1,36
Deusto	1,19
Ciemat	1,02
Url	1,02

Ciemat	1,39
Cervantes	0,55
Upsa	0,55
Fundesco	0,55
Cicyt	0,28
Deusto	0,28
Impi	0,28
Unnet	0,28
Vhebron	0,28
Hrc	0,28

Deusto	0,73
Ciemat	0,73
Upsa	0,37
Cervantes	0,37
Cicyt	0,37
Unnet	0,18
Vhebron	0,18
Hrc	0,18
Fundesco	0,18
Impi	0,18

Upco	1,02
Upsa	0,51
Cicyt	0,34
Unnet	0,34
Upna	0,34
Cervantes	0,34
Impi	0,17
Fundesco	0,17
Vhebron	0,17
Hrc	0,17

2.5.3. Tamaño de los documentos.

El tamaño informático (Aguillo, 2000) considerado también como indicador cibermétrico de primer orden nos permite comparar el tamaño en bytes de las diferentes páginas Web y en principio a mayor tamaño mayor cantidad de contenidos se ofrecen. A diferencia de lo expresado por (Aguillo, 2000) nuestra medida no se encuentra sesgada por la presencia de los objetos multimedia, pues el tamaño es el tamaño real del contenido, prescindiendo de estos elementos multimedia.

Esta es una de las ventajas de trabajar con una herramienta de desarrollo propio, que nos ha permitido afinar en los datos obtenidos de forma mucho más precisa.

Los tamaños medios de las páginas (en bytes) han ido evolucionando hacia

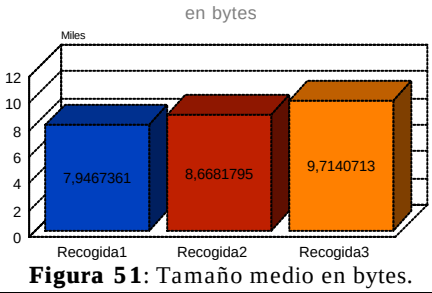


Figura 51: Tamaño medio en bytes.

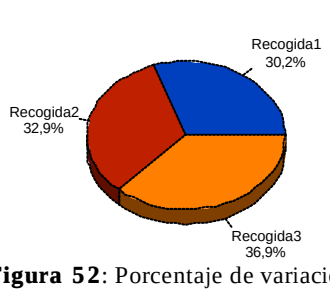


Figura 52: Porcentaje de variación del tamaño medio.

una media mayor según avanzamos en las recogidas, pero con un aumento pequeño de una recogida a la otra.

Es preciso indicar la enorme desviación existente en los tamaños de las páginas, que fluctúa entre 10708 y 13350 bytes, expresando una enorme variación en el tamaño.

Con respecto al tamaño medio, en bytes, de cada uno de los dominios, podemos observar que básicamente el mismo aumenta de una recogida a otra en una proporción moderada, excepto algún caso extremo, como el dominio *cicyt* que hay una considerable reducción de la primera recogida a la segunda o el caso de la *upsa*, que desciende en todas las recogidas.

Dominio	Re1	Re2	Re3
cervantes	8213,17	8852,55	9108,66
cicyt	17656,91	6113,79	6245,86
ciemat	7284,24	7581,02	8617,36
deusto	5710,46	7767,14	13858,11
esa	6148,3	6272,14	5996,47
fundesco	6959,35	5564,55	4333,73
hrc	0	12391,44	13347,82
impi	7540,85	8198,71	7170,77
uam	9173,31	12675,88	14310,61
uc3m	5393,02	8490,07	8724,28
ugr	8491,74	10462,85	10601,55
ujaen	6959,13	6769,67	7788,74
uji	8150,16	8179,35	10301,1
um	10462,95	9096,73	9850,43
unex	6075,8	7463,12	8000,17
unican	11363,31	12271,74	12431,42
unnet	7791,28	7797,9	7083,41
upco	6449,14	7213,21	6987,34
upna	7004,39	7132,03	18160,36
upsa	14255,67	12421,46	9273,46
url	8278,5	8351,8	8761,33
urv	4706,93	5444,1	7240,13
us	7906,8	11128,99	11021,67
usal	7154,46	7684,28	8885,84
uv	13935,02	14892,33	17217,7

uva	8569,12	9634,87	10329,46
vhebron	2927,89	4189,14	6632,14

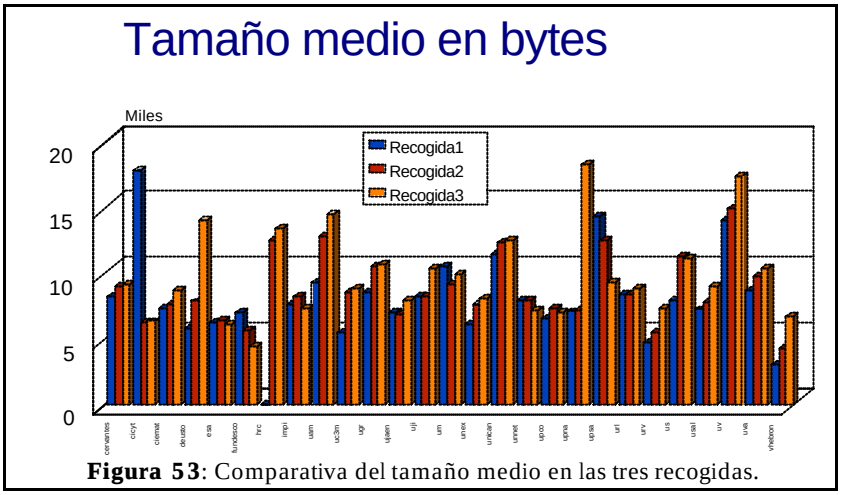


Figura 53: Comparativa del tamaño medio en las tres recogidas.

Parece evidente, que con la evolución en el lenguaje HTML y sobre todo con el empleo de editores HTML más sofisticados, que añaden código propio de gestión, el tamaño debe aumentar progresivamente.

En todas las recogidas el tamaño medio es superior al ofrecido en algunos trabajos anteriores (Bray, 1996) (Almind, 1997), que dan una media de 6,5 Kb y 6 Kb, aunque la desviación de nuestra investigación se mantiene bastante por debajo de los datos obtenidos por Bray (Bray, 1996).

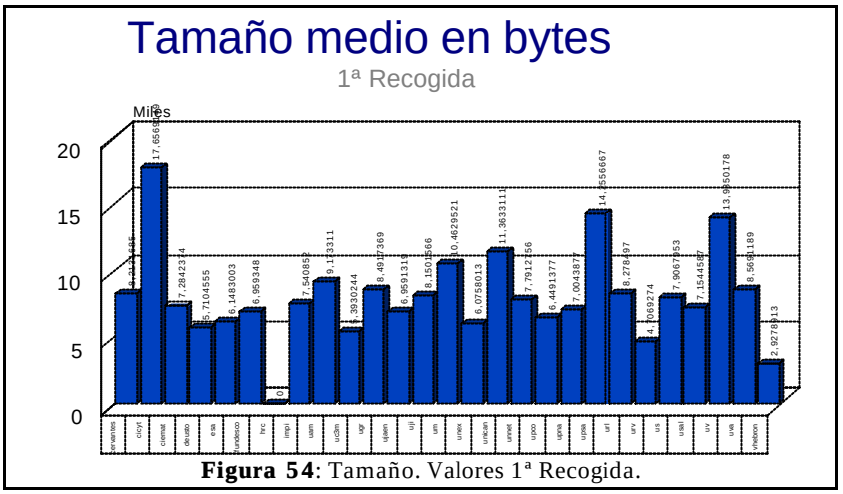


Figura 54: Tamaño. Valores 1ª Recogida.

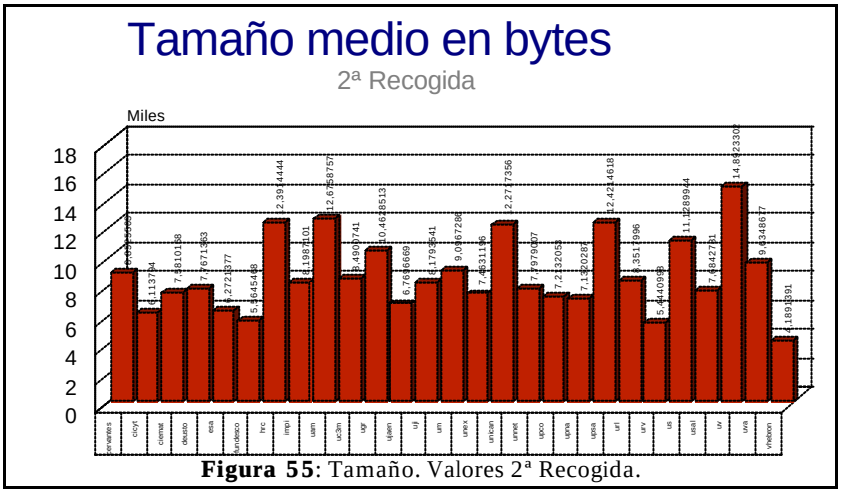


Figura 55: Tamaño. Valores 2ª Recogida.

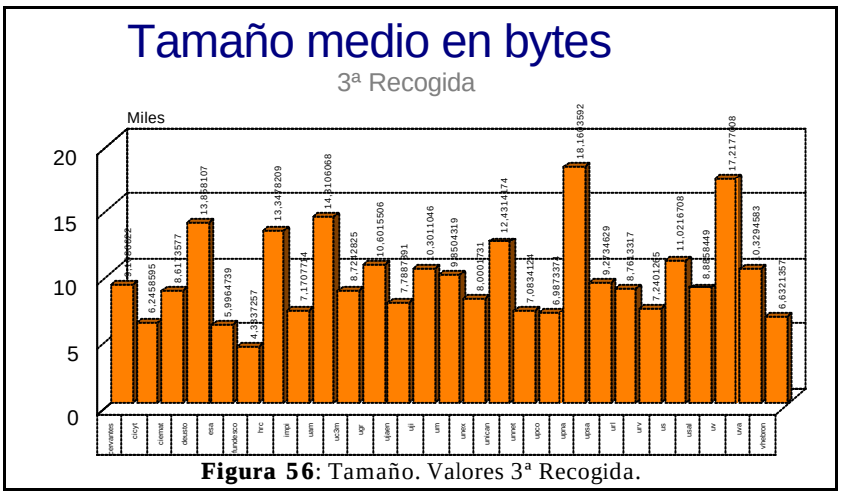


Figura 56: Tamaño. Valores 3ª Recogida.

2.6. ¿Cómo están conectados los dominios españoles?.

El análisis de las características que nos ofrecen los enlaces, como parte importante del hipertexto y por tanto del Web, es una parte esencial para conocer adecuadamente los diferentes dominios o sedes que estamos analizando y que está relacionado con la información y con los flujos de información que se originan en el Web, porque como dice (Alonso, 1999) los enlaces le confieren un carácter particular que lo diferencia del resto de los documentos y nos permite abordar la recuperación de información en el Web, aspecto fundamental en el tratamiento documental, como se recoge en (Figuerola, 1998).

2.6.1. Análisis hipertextual. El número de enlaces.

Hay que aclarar previamente un aspecto terminológico referido al número de enlaces. Algunos autores (Leydesdorff, 1999) hablan del concepto de popularidad, entendido como el número de visitas que recibe una sede determinada y no lo diferencian del número de enlaces. Claramente hay que distinguir entre la visita a una página y el enlace a una página y por supuesto son dos cosas diferentes. El concepto de popularidad ni siquiera se debe entender como un indicador.

La evolución correspondiente al número de enlaces (eliminados los autoenlaces de páginas) sigue la misma evolución de los dos apartados anteriores,

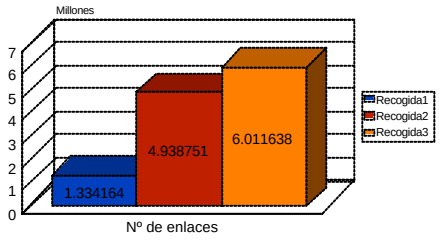


Figura 57: Número de enlaces en cada recogida.

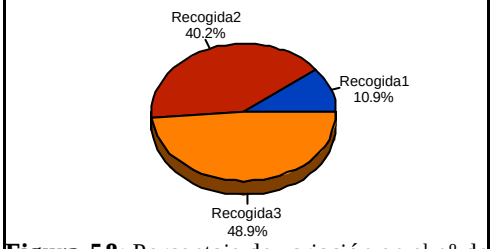


Figura 58: Porcentaje de variación en el nº de enlaces.

aunque el aumento entre la segunda y tercera recogida en este caso está un poco más acentuada.

Si analizamos los autoenlaces de nodo para el global de los dominios los datos son:

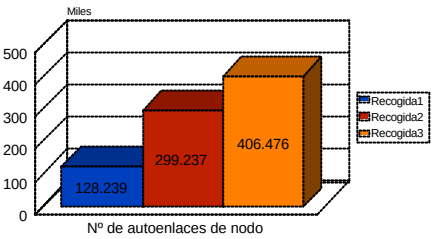


Figura 59: Número de autoenlaces.

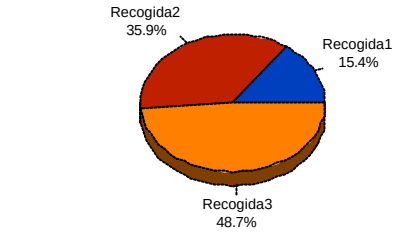


Figura 60: Porcentaje de variación en el nº de autoenlaces.

Básicamente la evolución es muy similar a la de los enlaces a otros nodos del mismo dominio, con una mayor evolución en la tercera recogida.

Estos datos globales vistos podemos precisarlos más si pasamos a la enumeración de determinado conjunto de índices o indicadores que pueden completar los datos globales.

2.6.2. Densidad hipertextual.

El término de densidad ha sido utilizado por diferentes autores y con diferente consideración, y uno de los primeros trabajos en considerar la densidad de los enlaces fue (Khan, 1998), pensada para su utilización en recuperación de la información.

Si seguimos a (Aguillo, 2000) esta medida sería la media aritmética en el número de enlaces que tiene cada uno de los dominios o sedes. Esta medida puede estar sesgada en el caso de tener una varianza elevada o unos rangos de mínimo y máximo amplios.

Los datos de la densidad para cada dominio analizado serían los siguientes:

Dominio	R1	R2	R3
cervantes	8,064644455	12,82007291	15,26172126
cicyt	2,042553191	6,136448598	8,393389501
ciemat	23,82149591	23,26540012	30,24174174
deusto	2,611650485	6,917737789	48,12173038

esa	18,15151515	13,82280502	15,04424937
fundesco	8,408284024	6,337962963	5,570412518
hrc	15,06666667	10,92307692	11,65492958
impi	4,161434978	4,177130045	4,387628866
uam	18,39259987	21,28928162	24,5583268
uc3m	10,93119196	14,96301826	15,65018879
ugr	9,23315047	8,522301192	9,827350236
ujaen	8,041814947	8,215304799	9,515312916
uji	13,93272311	10,9457805	13,63033615
um	8,197792869	8,446037902	11,76763346
unex	8,402095866	8,230318691	7,292714378
unican	13,86468595	13,29003654	14,23497322
unnet	18,42613636	18,21637427	14,92592593
upco	13,95781638	8,351800554	9,739981361
upna	11,53399258	21,19708423	37,43866944
upsa	1,994818653	4,274725275	4,265486726
url	11,67336908	9,632299638	11,6587856
urv	11,45746586	25,80890561	37,7894413
us	4,983804232	21,01529957	12,51651518
usal	5,668542744	6,724383724	7,774847307
uv	13,44777076	10,64964666	13,06378853
uva	10,14473171	9,368048622	8,739678741
vhebron	4,838659794	5,504950495	6,461295419

En general la tendencia es a aumentar esta densidad en cada nueva recogida, pero en la mayor parte de los casos es un aumento muy moderado. Algunos dominios descienden su densidad y también de forma moderada. Según estos datos la densidad de cada dominio, aún sufriendo variaciones, se mantiene en unos niveles bastante similares.

Ello nos indica que el número de enlaces medio de los dominios, entre las diferentes recogidas, es bastante estable, salvo alguna excepción. Lógicamente con esta medida no podemos valorar si los cambios que se han producido en los enlaces son importantes o no. Sabemos que existen modificaciones moderadas en cuanto a la densidad, pero para valorar el nivel de permanencia y volatilidad debemos acudir a otros factores.

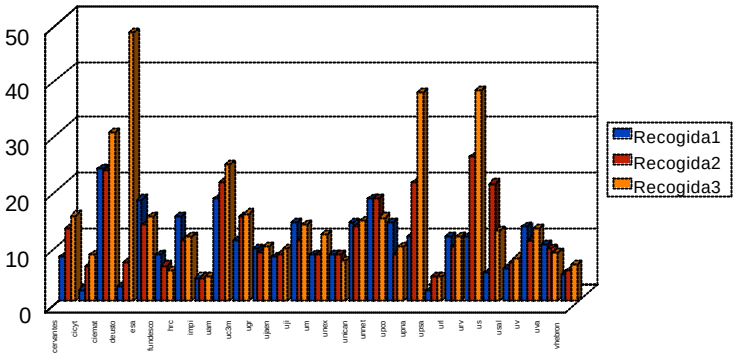


Figura 61: Variación en la densidad de enlaces en las tres recogidas.

Existen diferencias en el valor medio entre los diferentes dominios con un margen bastante amplio que indica una gran disparidad entre los diferentes dominios.

El término densidad empleado por (Almind, 1997) hace referencia al ratio tamaño de la página por el nº de enlaces. Esta medida puede ser válida para analizar cada nodo de una recogida a otra, pero si queremos analizar el dominio en su globalidad, nos obligaría a trabajar con el tamaño medio de cada página del dominio (que tiene una gran desviación) y con el número de enlaces medio de cada dominio, con lo que la validez de la medida se ve fuertemente influenciada y sesgada. Por ello creemos más correcto el empleo de la medida propuesta por (Aguillo, 2000).

2.6.3. Índice de desarrollo hipertextual.

La medida de la densidad puede tener algunos sesgos que pueden limitar la validez de los datos. Diversos autores han propuesto medidas que eliminan dichos sesgos, y en el que podemos valorar la evolución real del sistema hipertexto.

(Arellano, 1999) emplea el índice de hipertextualidad $\frac{Enlaces\ diferentes}{N^o\ de\ paginas}$ (sin distinguir el tipo de enlace) para conocer el grado de desarrollo hipertextual de cada sitio Web. (Rodriguez, 1997) emplea la misma fórmula pero no para conocer este desarrollo sino para conocer el impacto de los dominios. Para el cálculo del factor de impacto existen otros sistemas más fiables para realizar este tipo de comparativas como veremos más adelante.

(Ellis, 1994) y (Parunak, 1989) sugieren que una buena medida en este sentido es $\frac{n^o\ de\ nodos}{n^o\ de\ enlaces}$ (sin contabilizar los autoenlaces de nodo) que nos permite caracterizar la topología hipertexto, de tal forma que lo valores más bajos indican un mejor índice de hipertextualidad y por lo tanto se encuentran mejor conectados de forma global. Nosotros nos inclinamos por utilizar esta última corriente y son los datos que mostramos a continuación.

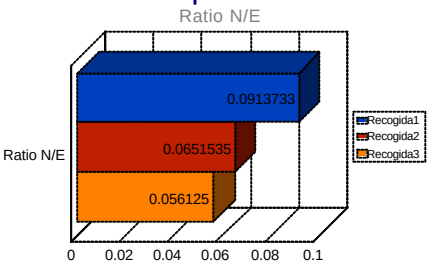


Figura 62: Valor del índice de hipertextualidad.

Como podemos observar, de forma global, este índice mejora en todas las recojidas y sobre todo de la primera recogida a la segunda.

El número de enlaces y de nodos aumentó también de forma considerable entre estas recogidas y este dato nos indica que fue un gran crecimiento, pero mejorando y poniendo mayor énfasis en la mejora de la estructura hipertextual de las diferentes sedes analizadas.

El ratio N/E individualizado para cada uno de los dominios es el siguiente:

Dominio	N/E-R1
ciemat	0,042
uam	0,054
unnet	0,054
esa	0,055

Dominio	N/E-R2
urv	0,039
ciemat	0,043
upna	0,047
uam	0,047

Dominio	N/E-R3
deusto	0,021
urv	0,026
upna	0,027
ciemat	0,033

hrc	0,066
upco	0,072
uji	0,072
unican	0,072
uv	0,074
url	0,086
upna	0,087
urv	0,087
uc3m	0,091
uva	0,099
ugr	0,108
fundesco	0,119
unex	0,119
um	0,122
ujaen	0,124
cervantes	0,124
usal	0,176
us	0,201
vhebron	0,207
impi	0,24
deusto	0,383
cicyt	0,49
upsa	0,501

us	0,048
unnet	0,055
uc3m	0,067
esa	0,072
unican	0,075
cervantes	0,078
uji	0,091
hrc	0,092
uv	0,094
url	0,104
uva	0,107
ugr	0,117
um	0,118
upco	0,12
ujaen	0,122
unex	0,122
deusto	0,145
usal	0,149
fundesco	0,158
cicyt	0,163
vhebron	0,182
upsa	0,234
impi	0,239

uam	0,041
uc3m	0,064
esa	0,066
cervantes	0,066
unnet	0,067
unican	0,07
uji	0,073
uv	0,077
us	0,08
um	0,085
hrc	0,086
url	0,086
ugr	0,102
upco	0,103
ujaen	0,105
uva	0,114
cicyt	0,119
usal	0,129
unex	0,137
vhebron	0,155
fundesco	0,18
impi	0,228
upsa	0,234

2.6.4. Índice de Endogamia.

Este índice propuesto por (Aguillo, 2000) trata de analizar la calidad del sistema hipertexto valorando los enlaces internos y ofreciendo una buena medida de si los dominios básicamente se referencian ellos mismos o si por el contrario tienen un nivel de conexión con otros dominios mejor. Para ello propone calcular $\frac{n^o\,de\,enlaces\,internos}{n^o\,de\,enlaces\,totales}$ de tal forma que cuanto menor sea este valor, menos endogámicos serán los dominios, es decir sus enlaces y por lo tanto los flujos de información de la sede no son a la misma página que posee el enlace, sino a páginas diferentes de la que enlaza, enriqueciendo el valor del propio enlace.

(Arellano, 1999) define este mismo índice pero lo denomina como índice de interconexión.

Los valores obtenidos para nuestros dominios son los siguientes:

Dominio	Endog-R1
Esa	0,01
Upsa	0,02
Uva	0,03
Cicyt	0,03
Urv	0,03
Impi	0,03
Vhebron	0,03
Uji	0,04
Deusto	0,04
Fundesco	0,04
Uc3m	0,06
Ugr	0,06
Unex	0,06
Url	0,08
Upco	0,09
Uv	0,09
Ciemat	0,1
Uam	0,11
Upna	0,11
Usal	0,11
Cervantes	0,12
Ujaen	0,13
Unnet	0,15
Us	0,16
Unican	0,18
Um	0,18
Hrc	0,2

Dominio	Endog-R2
Cicyt	0,01
Us	0,02
Impi	0,03
Upna	0,03
Urv	0,04
Uji	0,05
Uva	0,05
Esa	0,05
Fundesco	0,05
Uc3m	0,06
Ugr	0,06
Vhebron	0,06
Cervantes	0,06
Uam	0,07
Unex	0,07
Url	0,09
Usal	0,09
Deusto	0,09
Uv	0,09
Um	0,09
Ciemat	0,1
Ujaen	0,11
Upco	0,13
Unican	0,14
Unnet	0,15
Hrc	0,21
Upsa	0,22

Dominio	Endog-R3
Cicyt	0,01
Esa	0,03
Impi	0,03
Upna	0,04
Cervantes	0,04
Uji	0,05
Fundesco	0,05
Us	0,05
Urv	0,05
Uc3m	0,05
Uam	0,06
Unex	0,06
Uva	0,06
Um	0,06
Ugr	0,07
Deusto	0,07
Vhebron	0,08
Url	0,08
Usal	0,08
Uv	0,08
Unican	0,11
Ujaen	0,11
Ciemat	0,11
Upco	0,13
Upsa	0,14
Hrc	0,15
Unnet	0,16

2.7. Factor de Impacto Web (WIF).

El concepto de Factor de Impacto fue introducido por (Garfield, 1976) y sirve para medir la relación entre las citas recibidas en un determinado año, por los

trabajos publicados en una revista durante los dos años anteriores, y el total de artículos publicados en ella durante esos dos años anteriores (Sancho, 1990).

El WIF, índice propuesto por (Ingwersen, 1998) y adoptado por (Smith, 1999) trata de aplicar los mismos criterios que para el cálculo del factor de impacto en las publicaciones tradicionales, aunque las páginas Web son diferentes de los artículos de revista (Smith, 1999) y se precisa redefinir el cálculo de dicho indicador. (Cronin, 1996) dice que el Web es un buen sistema para publicar ciencia y por ello se puede aplicar el cálculo de un factor de impacto.

Se define como el cociente entre el número de enlaces externos recibidos por una sede y el tamaño de dicha sede, expresado por el número de páginas, $\frac{n^o\ enlaces que apun\tan fuera de la sede}{n^o\ de\ paginas totales}$ y nos ofrece una buena medida de qué dominios tienen un factor de impacto más elevado dentro de todos los dominios analizados significando que son los dominios más enlazados (citados) desde los otros dominios objeto del estudio. (Smith, 1999b) lo emplea para el análisis del Web australiano, utilizando Altavista para la recogida de datos. Una de sus conclusiones es que el WIF es una medida útil para ver la influencia de las sedes en su entorno.

Algunos autores han comentado que el cálculo del WIF tiene algunos sesgos, debidos al empleo de los buscadores de información como herramienta para poder realizar estos cálculos, pues la mayor parte de los estudios cibermétricos existentes se basan en el empleo de Altavista. Los sesgos que se le atribuyen son el de tener comportamientos irregulares en la presentación de los datos de tipo numérico, y con un comportamiento sospechoso respecto a la aplicación de los operadores booleanos (Notess, 2000).

(Snyder, 1999) indica que estos sesgos de los buscadores de información se centran en la inconsistencia del número de enlaces que devuelven, respecto a la consulta realizada y por ello finalmente el propio (Aguillo, 2000) propone la utilización de herramientas que obtengan los datos de forma directa.

Nuestro trabajo precisamente emplea una herramienta de recogida de este tipo, pues algunos de los problemas ya fueron planteados por nosotros en (Alonso,

1997) y por ello el WIF obtenido por nosotros carece de los sesgos planteados con anterioridad.

Los resultados para las tres recogidas son los siguientes:

Dominio	WebIF
uji	0,01925
uv	0,00564
uam	0,00519
ugr	0,00388
cicyt	0,00373
unican	0,00226
uva	0,00226
us	0,00211
um	0,0018
uc3m	0,00157
usal	0,0014
unex	0,00097
ujaen	0,00091
urv	0,00084
upna	0,00069
esa	0,00062
deusto	0,00056
fundesco	0,00055
url	0,0005
upco	0,00045
cervantes	0,00043
ciemat	0,00042
impi	0,00026
unnet	0,00025
upsa	0,00013
vhebron	0,0001
hrc	0,00003

Dominio	WebIF
uji	0,00531
us	0,00405
uv	0,00315
uam	0,00291
esa	0,00256
ugr	0,00232
cicyt	0,00161
uva	0,00152
uc3m	0,00135
unican	0,00135
um	0,00132
usal	0,00131
unex	0,00073
ujaen	0,00066
urv	0,00058
upna	0,00056
deusto	0,00041
url	0,00039
upco	0,00036
fundesco	0,00035
cervantes	0,00034
ciemat	0,00027
unnet	0,00018
upsa	0,00015
impi	0,0001
hrc	0,00007
vhebron	0,00004

Dominio	WebIF
uji	0,00526
us	0,00382
uam	0,0031
uv	0,003
esa	0,00269
ugr	0,00226
cicyt	0,00173
uva	0,0016
uc3m	0,00145
um	0,00133
usal	0,00129
unican	0,00122
unex	0,0009
ujaen	0,00069
urv	0,00065
upna	0,00061
deusto	0,00049
cervantes	0,00047
upco	0,00041
url	0,00038
fundesco	0,00033
ciemat	0,00027
upsa	0,00023
unnet	0,00023
impi	0,00008
hrc	0,00007
vhebron	0,00007

Siguiendo el mismo procedimiento, podemos calcular el WIF para otros dominios enlazados por los dominios objeto del estudio, ofreciendo una idea de los dominios que poseen una mayor factor de impacto, en cada una de las recogidas.

Dominio	WebIF-R1
elpais	0,04891
ole	0,01791
rediris	0,0119
cica	0,01038
arrakis	0,00998
mec	0,0081
upv	0,0079
ctv	0,00742
csic	0,0072
gva	0,00597
ucm	0,00595
recoletos	0,00574
redestb	0,00543
teknoland	0,0047
mcu	0,00463
el-mundo	0,00445
uniovi	0,0043
boe	0,00406
upc	0,00389
ub	0,00358
abc	0,00355
unizar	0,00331
upm	0,00308

Dominio	WebIF-R2
cica	0,02513
elpais	0,01432
rediris	0,00604
arrakis	0,00542
csic	0,00538
mec	0,00465
upm	0,00452
ucm	0,00389
ctv	0,00386
upv	0,00375
unizar	0,00356
ole	0,00341
uniovi	0,003
gva	0,00281
upc	0,0028
ub	0,00272
redestb	0,00263
ua	0,00258
mcu	0,00256
el-mundo	0,00238
cbuc	0,00224
recoletos	0,002
servicom	0,00168

Dominio	WebIF-R3
elpais	0,01416
mec	0,00577
rediris	0,00563
csic	0,00511
arrakis	0,00498
ucm	0,00437
ctv	0,00357
upv	0,00355
ole	0,00318
upm	0,00309
upc	0,00302
gva	0,00297
mcu	0,00288
ub	0,00287
cbuc	0,00285
uniovi	0,00275
cica	0,0027
unizar	0,00249
redestb	0,00234
el-mundo	0,00218
recoletos	0,00178
uab	0,00162
boe	0,00161

2.8. Visibilidad.

Este término empleado por (Bray, 1996), que lo definió como la característica que nos indica la sede que más enlaces recibe, ha sido redefinido por (Aguillo, 2000) pero normalizando su valor. La propuesta de (Aguillo, 2000) pretende eliminar los sesgos que puede tener el cálculo del WIF y comentados con anterioridad.

Se ha redefinido la visibilidad por parte de (Aguillo, 2000) como $\frac{n^o \text{ de paginas que ci tan la sede}}{n^o \text{ total de paginas de la sede evaluada}}$ consiguiendo obtener los dominios que más

enlaces reciben pero de forma normalizada, permitiendo establecer la comparación entre los diferentes dominios analizados y que a continuación ordenamos en orden decreciente de visibilidad.

Dominio	Visibili.-R1
Cicyt	0,968
Uji	0,517
Esa	0,206
Uam	0,202
Um	0,184
Unnet	0,176
Upna	0,102
Ujaen	0,097
Deusto	0,094
Hrc	0,088
Upsa	0,082
Fundesco	0,078
Impi	0,069
Upco	0,066
Ugr	0,057
Ciemat	0,032
Us	0,028
Uc3m	0,027
Url	0,026
Cervantes	0,025
Usal	0,024
Uv	0,023
Unex	0,021
Unican	0,019
Uva	0,011
Urv	0,01
Vhebron	0,001

Dominio	Visibili.-R2
Cicyt	0,483
Unnet	0,345
Hrc	0,263
Fundesco	0,175
Upsa	0,134
Deusto	0,112
Upco	0,106
Uji	0,101
Upna	0,097
Ujaen	0,091
Impi	0,073
Url	0,049
Ciemat	0,049
Usal	0,04
Unex	0,039
Ugr	0,039
Esa	0,032
Um	0,029
Uam	0,023
Cervantes	0,023
Uv	0,022
Unican	0,021
Uva	0,019
Uc3m	0,018
Us	0,017
Vhebron	0,014
Urv	0,006

Dominio	Visibili.-R3
Upna	0,404
Unnet	0,347
Cicyt	0,329
Upsa	0,172
Hrc	0,169
Fundesco	0,15
Upco	0,095
Ujaen	0,063
Url	0,056
Uji	0,054
Ciemat	0,054
Impi	0,053
Ugr	0,039
Usal	0,038
Um	0,035
Unex	0,029
Deusto	0,025
Esa	0,025
Cervantes	0,024
Uva	0,022
Uc3m	0,022
Us	0,02
Unican	0,019
Uam	0,019
Uv	0,016
Vhebron	0,015
Urv	0,004

2.9. Análisis de Citas.

(Sandison, 1989) indica que una cita no es un conjunto de datos bibliográficos al final de un documento en forma de notas final o de notas a pie de página, etc. Una cita es la representación de una decisión realizada por un autor

que desea mostrar la relación entre el documento que él está escribiendo y el trabajo de otro autor.

Según (Egghe, 1990) la existencia de un documento citado en una lista de referencias indica que existe una relación entre el documento citado y el que cita, desde el punto de vista del autor.

Similarmente, (Smith, 1981) indica que la relación entre el documento citado y el que cita se representa mediante una cita. La naturaleza de esta relación es difícil de caracterizar, aunque Smith (1981) apunta las siguientes razones para citar un documento:

1. Pagar tributo a los predecesores.

2. Dar crédito a trabajos relacionados (tributo a los contemporáneos)

3. Identificación de metodología y equipos.

4. Ofrecer antecedentes a los lectores.

5. Corrección de nuestros propios trabajos.

6. Corrección del trabajo de otros.

7. Crítica de trabajos previos.

8. Comprobar reivindicaciones.

9. Alertar de próximos trabajos.

10. Ofrecer pistas de trabajos poco diseminados, pobremente indexados o que no han sido citados.

11. Identificar publicaciones originales en las cuales una idea o concepto se discutió.

12. Identificar publicaciones originales u otros trabajos que describen un concepto o término eponímico.

13. Negar el trabajo o las ideas de otros.

14. Disputar las reclamaciones prioritarias de otros (negación de tributo).

Garfield (1978) describe el análisis de citas como una herramienta analítica que utiliza las citas y referencias de los documentos científicos. Le Pair (1988) dice que el análisis de citas es probablemente una buena herramienta de evaluación para los campos científicos que utilizan la revista como canal de comunicación. Lancaster (1991) indica la existencia de un importante grupo de estudios bibliométricos relacionados con las citas de los autores. Indica que el análisis de citas debe responder a aspectos como: quien cita a quien, que revista es citada por otra, qué campos temáticos son más citados en la literatura de una disciplina específica, etc.

Algunas de las aplicaciones del análisis de citas puede ser las siguientes:

Según Garfield (1979) la simplificación de los procesos investigadores y el aumento en los resultados de la investigación son posibles mediante la utilización de métodos investigadores basados en el análisis de citas. Lord (1984) indica que puede emplearse el análisis de citas para contar citas que permitan evaluar las publicaciones científicas, para conocer la bibliografía citada por dos documentos o para un análisis de cocitas, que nos permiten estudiar el desarrollo de la ciencia en un campo específico.

Egghe (1990) indica tres aplicaciones principales: "evaluación cuantitativa y cualitativa de los científicos, de las publicaciones y de los organismos investigadores; modelar el desarrollo histórico de la ciencia y de la tecnología; y para búsqueda y recuperación de la información".

Centrándonos en el análisis de cocitas, Small(1973) lo describe como "la frecuencia con la que dos documentos son citados juntos". Cawkell (1976) presenta una definición similar al de los indicadores de similaridad y demuestra que para la realización de un análisis de cocitas podemos trabajar con una matriz de citas, de forma similar al esquema utilizado en el tratamiento de grafos.

(Larson, 1996) en su trabajo ya hace referencia al análisis de cocitas y al Web indicando que sería útil para ofrecer una idea de la estructura intelectual, como indicó (Bayer, 1990). Larson hace referencia a los trabajos de (McCain, 1990) y (McCain, 1991) indicando algunos de los campos y las técnicas en las que

se podía aplicar. Larson también hace referencia al empleo de las matrices para el tratamiento adecuado de las cocitas.

(Cui, 1999) aplica las técnicas del análisis de citas a sedes Web especializadas en medicina, para medir las webs de medicina más citadas que les permitió conocer las instituciones médicas más prestigiosas desde este punto de vista.

Para medir la fuerza de la cocita se han presentado varias propuestas, pareciéndonos la más adecuada la de Garfield (1980) que explica la fuerza de la cocita (o porcentaje de solapamiento) mediante la siguiente fórmula: $S = \frac{Cocitas de A + B}{(Total de citas de A + B) - (Cocitas de A + B)}$

Si consideramos que los enlaces de las páginas Web, pueden realizar las funciones de las citas y que podemos utilizar las matrices para realizar la representación de las citas recibidas por cada uno de los dominios, puede resultar interesante realizar un análisis de cocitas clásico, con el fin de obtener que dominios son los que están más relacionados y con una mayor fuerza.

Siguiendo a (Osareh, 1996) podemos definir la fuerza de la cocita como $S = \frac{Cocitas de A + B}{(Total de citas de A + B) - (Cocitas de A + B)}$. Aplicándolo a los dominios objeto de estudio obtenemos los siguientes resultados:

Dominios	Fuerza-R1
2 - 13	150
2 - 9	81,46
9 - 13	76,84
13 - 16	65,05
13 - 14	56,67
12 - 22	50,25
10 - 12	47,5
13 - 24	43,09
10 - 19	43
18 - 22	36,25
13 - 15	35,55
13 - 23	34,74

Dominios	Fuerza-R2
18 - 22	94,33
2 - 13	81,38
4 - 19	61,2
6 - 19	57,8
3 - 11	51,8
3 - 21	51,5
4 - 22	49,33
19 - 22	49,14
13 - 19	46,5
4 - 17	46,5
17 - 22	45
3 - 24	43,91

Dominios	Fuerza-R3
18 - 19	168,5
6 - 18	123
3 - 21	108
2 - 13	80,63
6 - 19	77,25
13 - 18	72,33
1 - 18	72,25
4 - 19	72,2
18 - 20	70,67
17 - 18	69,67
13 - 19	69,19
4 - 13	65,19

4 - 12	34,6	15 - 22	43,56	1 - 12	63,5
6 - 13	34,29	4 - 13	41,83	4 - 18	59,2
12 - 21	33	4 - 21	41,5	4 - 11	58,13
14 - 19	32,33	11 - 19	41,19	4 - 10	46,62
10 - 13	30,74	18 - 19	41,14	18 - 22	45,86
19 - 22	28,67	9 - 19	39,44	13 - 15	45,76
9 - 23	28,52	4 - 24	37,5	1 - 4	44,71
19 - 21	27,6	3 - 9	37,21	3 - 9	44,14
12 - 13	27,25	12 - 22	37,2	1 - 6	43,5
4 - 13	26,75	3 - 10	36,62	17 - 22	43,33
13 - 19	25,97	3 - 13	34,65	1 - 11	42,95
11 - 13	24,77	9 - 21	34,37	3 - 16	42,27
1 - 13	24,43	3 - 14	34,29	11 - 19	41,05
9 - 14	24,27	4 - 6	34	1 - 13	39,43
10 - 14	23,88	3 - 25	33,71	15 - 22	39,42
3 - 13	23,62	4 - 12	33,2	1 - 19	38,78
12 - 19	23,12	15 - 19	33,17	19 - 22	38,3
9 - 16	23	2 - 24	32,04	4 - 6	38,29
2 - 10	22,56	19 - 24	31,67	11 - 18	38,09
4 - 22	22,29	6 - 16	31,62	17 - 20	37,25
6 - 9	21,03	22 - 24	31,17	6 - 20	36,8
14 - 23	20,71	21 - 24	30,29	6 - 13	36,7
10 - 22	20,23	6 - 15	30,27	3 - 26	36,12
4 - 14	19,36	16 - 19	29,95	4 - 21	35,5
13 - 22	18,82	13 - 15	29,41	3 - 11	34,35
4 - 17	18,8	12 - 17	29	3 - 13	33,96
14 - 21	18,79	4 - 14	28,94	18 - 23	33,68
13 - 26	18,78	17 - 19	28,88	19 - 23	33,48

Aclaración: Hay que aclarar que en el campo dominios los números que aparecen hacen referencia a los dominios, ordenados según aparece en el apéndice de la página xiii.

En la primera recogida, los dominios con mayor relación son *Cicyt-Uji*, seguido por *Cicit-Uam*, en la segunda recogida son *Upco-Urv*, seguidos de *Cicyt-Uji* y en la tercera *Upco-Upna*, seguidos de *Fundesco-Upco*.

2.10. Validez de los enlaces.

Un aspecto que puede resultar interesante es el de la permanencia de las páginas, tanto de las páginas con enlaces válidos como las páginas que tienen enlaces erróneos y relacionado con este último apartado el porcentaje de corrección de los enlaces erróneos, que nos dan una medida de la frecuencia de actualización y de ml mantenimiento de las diferentes sedes objeto del estudio.

Comenzando por la permanencia de las páginas, es decir, los enlaces a una página que en una de las recogidas eran válidos y que en la siguiente recogida continúan siendo válidos y por lo tanto sus enlaces siguen siendo los mismos y siguen siendo correctos, los datos para cada dominio son los siguientes:

Dominio	Permanece1-2	Permanece2-3
Cervantes	88,36	12,8
Cicyt	0	95,79
Ciemat	89,94	78,47
Deusto	0,42	77,38
Esa	49,86	93,57
Fundesco	30,06	69,6
Hrc	77,78	79,12
Impi	91,25	55,38
Uam	94,6	24,26
Uc3m	69,45	22,43
Ugr	23,63	31,14
Ujaen	76,6	78
Uji	76,15	75,19
Um	98,81	43,39
Unex	63,85	53,37
Unican	84,34	73,45
Unnet	94,32	100
Upco	0	81,9
Upna	20,27	0,53
Upsa	62,18	0
Url	45,17	54,89
Urv	73,77	58,84
Us	23,41	10,86
Usal	55,11	79,79

Uv	69,46	74,39
Uva	32,03	65,89
Vhebron	33,76	0

Como podemos ver, en algunos de los dominios el porcentaje de permanencia es 0, y ello significa que los enlaces han cambiado en su totalidad y no existe ningún enlace que coincida. Esto puede deberse a que el nombre de las páginas enlazadas ha sido modificado o bien que la ubicación física de dichas páginas también ha cambiado o que las direcciones IP de los servidores han cambiado, modificándose en todos los casos la URL del enlace.

Podemos valorar el porcentaje global de enlaces que permanecen entre las diferentes recogidas, reflejándose en los siguientes gráficos.

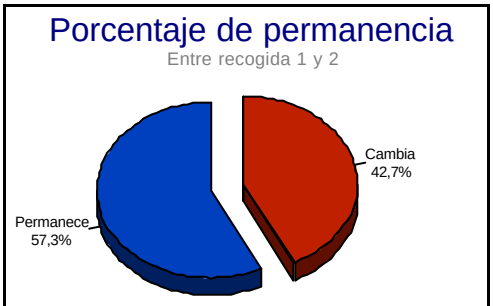


Figura 63: Porcentaje de permanencia entre las dos primeras recogidas.

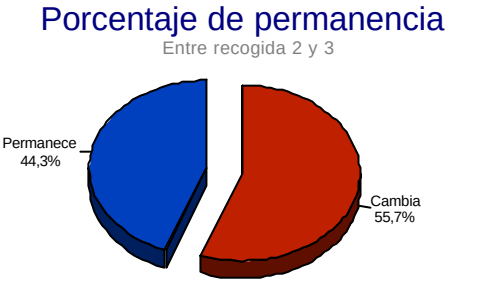


Figura 64: Porcentaje de permanencia entre las dos últimas recogidas.

El porcentaje de permanencia es más favorable en las dos primeras recogidas, que en las dos últimas, donde hay más cambios.

Respecto al porcentaje de enlaces que indicaban páginas no existentes (enlaces erróneos o muertos) que siguen manteniendo el mismo enlace erróneo los datos son los siguientes:

Dominio	Error1-2	Error2-3
Cervantes	72,72	5,45
Cicyt	0	57,14
Ciemat	99,07	99,95
Deusto	0	49,06
Esa	55,56	69,39

Fundesco	8,33	43,33
Hrc	100	75
Impi	0	50
Uam	98,6	17,85
Uc3m	16,08	32,14
Ugr	19,68	35,04
Ujaen	84,81	92,68
Uji	68,29	41,14
Um	54,78	68,96
Unex	38,26	41,93
Unican	75,44	72,04
Unnet	95,83	100
Upco	0	100
Upna	55,06	0
Upsa	76,92	0
Url	32,51	52,63
Urv	48,57	39,28
Us	14,06	52,97
Usal	60,71	60,07
Uv	92,32	69,18
Uva	26,51	72,91
Vhebron	86,27	0

Podemos realizar también una valoración global de las recogidas, ofreciendo los datos en los siguientes gráficos, donde podemos observar que en las

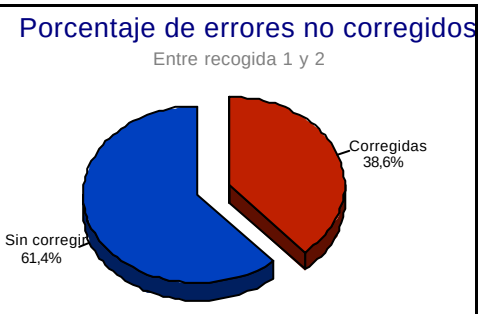


Figura 65: Porcentaje de errores no corregidos en las dos primeras recogidas.

Figura 66: Porcentaje de errores no corregidos en las dos últimas recogidas.

últimas recogidas aumenta la corrección de enlaces erróneos de la recogida anterior.

Respecto a las páginas con enlace erróneo en una recogida y que en la siguiente tienen ese mismo enlace para que ahora apunta a una página que sí existe y por lo tanto son enlaces que se han corregido:

Dominio	Corrige1-2	Corrige2-3
Cervantes	0	0
Cicyt	0	17,86
Ciemat	0,1	0
Deusto	0	0,94
Esa	0	0
Fundesco	0	10
Hrc	0	0
Impi	0	0
Uam	2,38	0,6
Uc3m	0,61	0,17
Ugr	10,53	1,29
Ujaen	1,27	0,81
Uji	0,81	0,75
Um	0,87	1,75
Unex	0,7	1,61
Unican	1,58	0,52
Unnet	0	0
Upco	0	0
Upna	0,63	0
Upsa	3,85	0
Url	0,49	1,5
Urv	3,36	1,31
Us	0,52	0,69
Usal	0,45	1,81
Uv	9,17	1,54
Uva	1,03	1,01
Vhebron	0	0

2.11. Diámetro Web.

Algunos autores (Aguillo, 2000) hablan del nivel de profundidad de la sede, como una indicación de la posible existencia de zonas que puedan permanecer

invisibles para el acceso, en el caso de existir una profundidad elevada. Sin embargo, esta medida, que ofrece una idea de lo difícil que puede llegar a ser alcanzar determinada información , no es realmente tan importante a la hora de valorar este aspecto, porque la profundidad tal y como la considera (Aguillo, 2000), se refiere al número de niveles existentes en el teórico árbol jerárquico que constituiría una sede Web. La medida que realmente valora este aspecto es el diámetro (Albert, 1999) y (Faloutsos, 1999) o distancia máxima para alcanzar un determinado documento, que no se relaciona directamente con el nivel de profundidad que exista.

Este aspecto se tratará más detenidamente en la tercera ley de exponenciación, en el capítulo 4, aunque a modo de avance, indicaremos que los datos obtenidos del análisis del grafo, en las condiciones indicadas anteriormente (grafo sobre 1000 páginas), son los siguientes:

Dominio	R1	R2	R3
Cervantes	13	12	11
Cicyt	3	16	11
Ciemat	7	6	6
Deusto	11	15	11
Esa	10	10	10
Fundesco	16	15	13
Hrc	5	7	8
Impi	15	15	9
Uam	11	13	11
Uc3m	47	9	11
Ugr	14	13	16
Ujaen	12	9	8
Uji	6	10	9
Um	8	8	7
Unex	15	16	15
Unican	7	8	10
Unnet	7	6	10
Upco	16	30	25
Upna	17	10	9
Upsa	10	11	12
Url	14	15	20

Urv	11	13	12
Us	10	14	14
Usal	9	12	12
Uv	7	8	21
Uva	9	11	10
Vhebron	19	27	16

Los dominios que poseen un diámetro elevado nos indican que poseen zonas de su estructura que son más difíciles de alcanzar. En general, todos los dominios mantienen unos diámetros reales bastante similares entre las diferentes recogidas, salvo algunas excepciones.

El diámetro medio en cada recogida es muy similar como podemos ver en los siguientes gráficos para cada una de las recogidas.

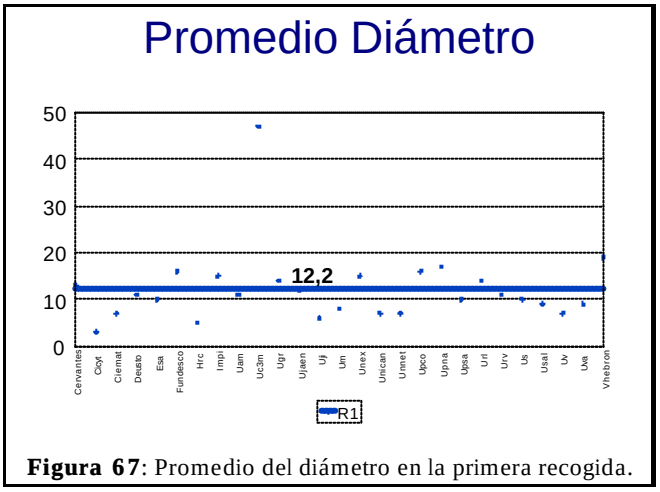


Figura 67: Promedio del diámetro en la primera recogida.

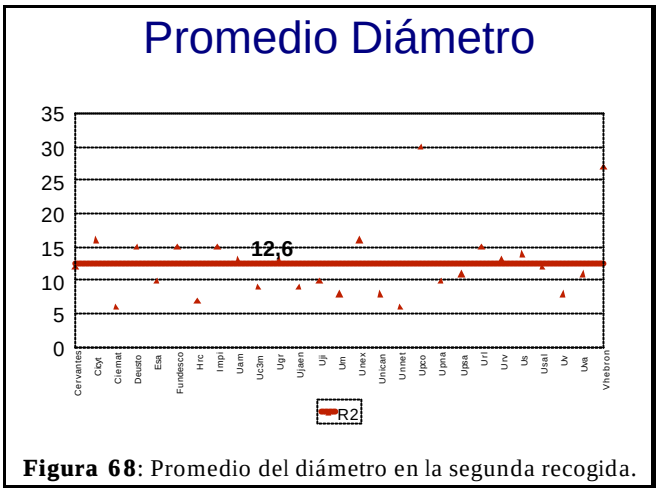


Figura 68: Promedio del diámetro en la segunda recogida.

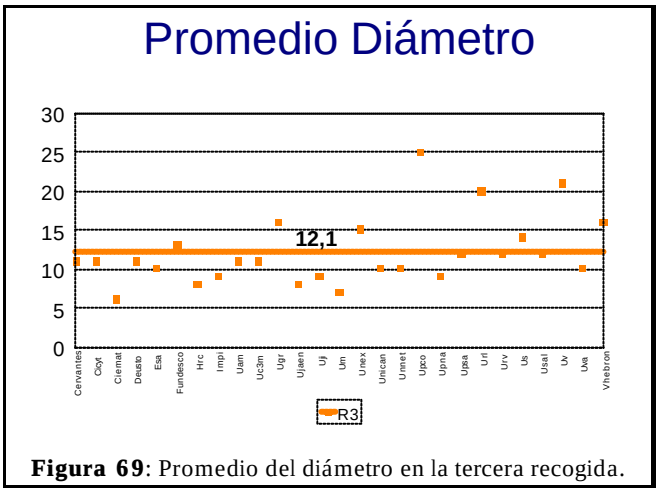


Figura 69: Promedio del diámetro en la tercera recogida.

A modo indicativo, ponemos a continuación la profundidad (número de niveles) del total de los datos de cada dominio y la profundidad para los primeros 1000 nodos:

Dominio	Total-R1	Total-R2	Total-R3	Par-R1	Par-R2	Par-R3
Cervantes	19	65	39	8	5	4
Cicyt	3	17	38	3	12	6
Ciemat	50	50	49	5	5	4
Deusto	8	15	25	8	9	7
Esa	7	121	475	7	5	5
Fundesco	11	10	10	11	9	9
Hrc	4	4	5	4	4	5
Impi	5	6	6	5	6	6
Uam	7	255	255	5	5	4
Uc3m	367	30	363	9	5	5
Ugr	36	71	258	7	5	4
Ujaen	24	17	17	8	5	4
Uji	8	104	625	4	4	4
Um	5	310	311	4	4	4
Unex	15	27	41	6	6	7
Unican	361	362	363	3	4	4
Unnet	4	5	6	4	5	6
Upco	9	19	15	9	12	9
Upna	12	13	6	12	6	5
Upsa	7	8	9	7	8	9
Url	19	18	23	10	9	12
Urv	76	73	115	6	6	5
Us	7	18	37	4	6	5
Usal	16	29	29	4	5	4
Uv	70	254	80	3	3	3
Uva	55	157	157	4	4	4
Vhebron	12	21	11	10	21	8

Como podemos ver, en muchos casos la profundidad para los 1000 primeros nodos está por debajo de la distancia o el diámetro real que tienen dichos dominios. Esto nos indica que la distancia para poder alcanzar esos documentos dentro del dominio es mayor que el número de niveles que lo forman y por ello es mucho más adecuado tratar con la distancia que con la profundidad.

2.12. Conclusiones.

Vamos a realizar un resumen de los aspectos más relevantes del análisis cuantitativo del Web.

1. Respecto a la evolución de los diferentes tipos de ficheros:

    a. Los ficheros de compresión tienden claramente al empleo del formato ZIP, que aumenta espectacularmente entre la primera y la segunda recogida, para compartir espacio junto al formato Z en la terecra recogida.

    b. Los ficheros gráficos reflejan que el formato más utilizado es el GIF, perdiendo algo de su cuota en la segunda recogida a favor del formato JPEG para volver a recuperar en la última recogida.

    c. En los ficheros de video el formato AVI y MPEG compartieron terreno en la primera recogida, para en la segunda el formato MPEG coger la mayor parte de la cuota y finalmente repartirse más o menos de forma similar la cuota junto al formato MOV.

    d. Los ficheros de sonido que partían con un poder casi absoluto del formato WAV, han visto como se repartía el poder para pasar a manos del formato AU, que en la última recogida tienen más del 50%.

    e. Los ficheros de texto tienden hacia un mayor empleo del formato PDF en detrimento del formato TXT. Destaca el aumento del formato TEX, que es un dato curioso, ya que lo normal es que los documentos realizados en TEX se ofrezcan en otro formato.

    f. El empleo de estilos aumenta en todas las recogidas de forma significativa.

    g. El empleo de elementos multimedia en los diferentes dominios se mantiene bastante estable entre las recogidas, con un ligerísimo aumento.

2. Respecto al empleo de determinadas etiquetas:

a. La etiqueta Title es mayoritariamente empleada.

b. Se emplea más la etiqueta FONT que la etiqueta H.

c. Las listas se encuentran en desuso frente al mayor empleo de la etiqueta TABLE.

d. Hay una escasa incidencia de la etiqueta MAP.

e. Se emplea más la opción BGCOLOR que la BACKGROUND, en la etiqueta Body.

f. Se aprecia un mayor empleo de Applets, mientras el empleo de scripts se mantiene bastante similar, con empleo de lenguaje Java de forma mayoritaria, aunque disminuyendo su porcentaje frente a otros lenguajes de script.

g. El empleo de frames desciende de forma importante.

h. Los formularios se emplean cada vez más, aunque cada vez menos para el envio de datos a través de correo electrónico.

3. El empleo de servidores Web está marcado por el empleo mayoritario del Servidor Apache, seguido de lejos por los servidores de Microsoft.

4. El estándar de exclusión de robots se emplea cada vez más, aunque el correcto empleo de dicho estándar empeora con cada una de las recogidas. El empleo de la etiqueta META pa excluir se emplea poco y no siempre de forma correcta.

5. El número de nodos y el de servidores aumenta en todas las recogidas, pero de forma más significativa entre la primera y la segunda recogida.

6. El tamaño medio en bytes de los diferentes documentos aumenta en todas las recogidas, aunque siempre con una enorme desviación en los datos.

7. El número de enlaces de los dominios aumentan de forma considerable entre la primera y la segunda recogida y en menor medida en la tercera.

8. La densidad hipertextual aumenta ligeramente en cada recogida.

9. El índice de desarrollo hipertextual mejora en todas las recogidas y sobre todo de la primera recogida a la segunda, fiel reflejo de lo comentados para el número de enlaces y nodos.

10. El factor de impacto Web y la visibilidad nos permiten conocer los dominios más enlazados.

11. Podemos aplicar un análisis de cocitas clásico a los diferentes dominios del estudio, obteniendo los dominios que mayores relaciones tienen entre sí.

3. Medidas Topológicas.

3.1. Introducción.

Una buena forma de analizar la evolución de los dominios Web, que se centra fundamentalmente en su naturaleza hipertextual, es calcular un conjunto de medidas que tienen en cuenta los enlaces que se producen entre los diferentes documentos que conforman el dominio correspondiente como justifican (Smeaton, 1995b), (Ellis, 1994), (Botafogo, 1992).

Otros estudios posteriores se han centrado en el estudio de la conectividad y la estructura topológica del Web como los de (Abraham, 1997), (Kleinberg, 1999) y (Wheeler, 1999) basándose en los trabajos anteriormente mencionados. Incluso (Kleinberg, 1999) relaciona este tipo de trabajos con los análisis de citas y con el factor de impacto.

Estas medidas se basan en la consideración del Web como un grafo (Botafogo, 1992), (Ellis, 1994), (Kleinberg, 1999b), (Hayes, 2000), (Broder, 2000) y la aplicación de diferentes técnicas propias de esta teoría, que analizaremos posteriormente.

Además estas medidas tienen un carácter interesante, propuesto por (Ellis, 1994), que es el de ser unas medidas que pueden considerarse como medidas de similaridad para los grafos. Esta consideración nos va a permitir además de obtener un valor único para cada uno de los grafos, poder utilizar ese valor único para realizar comparaciones entre los valores de diferentes dominios, indicándonos en el caso de coincidir dichos valores que ambos grafos (y por lo tanto el dominio) son muy similares desde el punto de vista de la característica que estemos midiendo en ese momento.

Cualquier medida de un grado de similaridad entre dos objetos debe realizarse al menos en dos pasos (Ellis, 1994):

1. En primer lugar hay que seleccionar los atributos de los objetos que van a estudiarse y que van a caracterizar dichos objetos.

2. Hay que seleccionar la medida o coeficiente adecuado, cuyo valor se deriva del análisis de los dos objetos implicados.

Para poder emplear adecuadamente este tipo de medidas de similaridad vamos a considerar los dominios Web como un grafo y aplicando esta técnica para su estudio, vamos a poder crear unos atributos adecuados y manejar unos coeficientes que nos permitan esta comparación.

3.2. Estudio del grafo.

En la consideración del Web como un grafo, los nodos se representan mediante las páginas html y los enlaces se representan mediante los bordes dirigidos. Diferentes estudios (Bharat, 1998) sugieren la existencia de varios cientos de millones de nodos en el grafo Web y con un crecimiento importante, y el número de enlaces alcanzaría varios billones (Kleinberg, 1999b). Algunos de los trabajos que han manejado el Web como un grafo han utilizado un volumen de información realmente importante con 200 millones de páginas y 1,5 billones de enlaces (Kumar, 1999b) mostrando las consistencia de los planteamientos y con la aplicación de algoritmos adecuados para el tratamiento de esta gran cantidad de información (Kleinberg, 1999b), (Dean, 1999).

El análisis de la estructura del grafo Web se ha empleado en ocasiones para mejorar la calidad de las búsquedas en el Web como en (Bharat, 1998b), (Brin, 1998), (Carriere, 1997), (Chakrabarti, 1998), (Kleinberg, 1999).

También se ha utilizado para clasificación de páginas Web en función de las materias de las páginas a las que apunta una página concreta como en (Chakrabarti, 1998b); para mostrar la información (Botafogo, 1991), (Pirolli, 1996), (Carriere, 1997); en minería del Web (Mendelzon, 1995), (Mendelzon, 1997), (Kumar, 1999), (Kumar, 1999b).

Algúnos autores han utilizado el Web como grafo para crear de forma automática hipertextos, partiendo de textos carentes de enlaces (Smeaton, 1992), (Smeaton, 1995), (Gollogley, 1997).

La estructura de enlaces del Web contiene también información sobre las diferentes comunidades Web que se pueden crear y que se reflejan mediante la topología del Web como apunta (Gibson, 1998) y también permite aplicar técnicas

de similaridad, basadas en los enlaces, para estructura y visualizar el Web (Chen, 1997), (Chakrabarti, 1999).

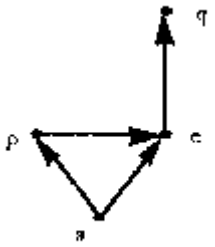
Aplicando la terminología de (Harary, 1969) en teoría de grafos al grafo Web, las páginas Web se denominarían nodos y los enlaces se denominarían como arcos o aristas.

La adecuada representación de un grafo, para su análisis automatizado (Ellis, 1994), puede realizarse mediante la creación de las llamadas matrices de adyacencia.

Para un grafo dirigido *G*, con *p* nodos, podemos definir la matriz de adyacencia *A*, formada por *p x p* elementos de la forma a_{ij}. Los valores que pueden tener los elementos de la matriz *A*, pueden ser:

1. Valor 1 si los dos nodos v_i y v_j están conectados mediante un enlace que va de v_i a v_j.
2. Valor 0 si no existe enlace desde v_i a v_j.
3. Valor nulo si i=j.

Un ejemplo de lo comentado con anterioridad se refleja en la siguiente matriz de adyacencia que representa al grafo adjunto, formado por cuatro nodos y que tiene una distribución de enlaces dirigidos en una dirección muy concreta. La matriz es el reflejo de dicha distribución.

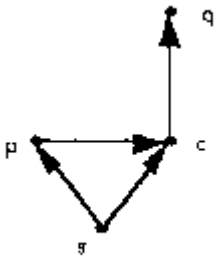


a_{ij}	a	b	c	d
a	-	1	1	0
b	0	-	1	0
c	0	0	-	1
d	0	0	0	-

De forma similar, podemos definir la matriz de distancia *D*, formada por *p x p* elementos de la forma d_{ij}, donde d_{ij} es la longitud (medida en enlaces) del camino más corto entre los nodos v_i y v_j si están conectados, valor cero si no están conectados y valor nulo si i=j.

Para la obtención de la matriz de distancia pueden aplicarse diferentes algoritmos aunque (Smeaton, 1995b) recomienda el algoritmo de Floyd recogido en (Sedgewick, 1990) y que representamos a continuación:

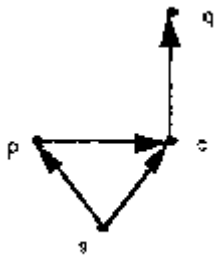
```
for k = 1:n
        for i = 1:n
                for j = 1:n
                        Y(i,j) = min (Y(i,j),Y(i,k) + Y(k,j));
                end
        end
end
```



d_{ij}	a	b	c	d
a	-	1	1	2
b	0	-	1	2
c	0	0	-	1
d	0	0	0	-

(Botafogo, 1992) introduce el concepto de Matriz de distancia convertida, que se puede representar mediante D'_k y se define por *p x p* elementos de la forma d'_{ijk}, donde d'_{ijk} es igual a d_{ij} (si d_{ij} ... 0) o tiene un valor que coincide con el número de nodos (si $d_{ij} = 0$).

Botafogo pretende indicar con ello que la distancia entre dos nodos no conectados no tiene un valor infinito como se puede interpretar del valor 0 y por ello sugieren que se sustituya por el valor obtenido de $\max(d_{ij})+1$, siendo $\max(d_{ij})$ el valor máximo que d_{ij} puede alcanzar, y la distancia máxima que se puede alcanzar será siempre $(n^o\, nodos - 1)$.



d'_{ij}	a	b	c	d
a	-	1	1	2
b	4	-	1	2
c	4	4	-	1
d	4	4	4	-

3.3. Índices aplicables.

Una vez representado el grafo de forma adecuada, y creadas las matrices correspondientes debemos pasar a analizar los posibles índices que pueden estar

a nuestro alcance para el estudio de dicho grafo (y por aplicación a nuestros dominios).

Los índices los podemos clasificar de la siguiente forma:

3.3.1. Índices de nodo.

Se valora la similaridad de los nodos entre los dominios. Al obtener valores normalizados nos permite comparar los valores entre dos recogidas permitiéndonos saber si son muy parecidos o no. Las medidas de grado de apertura, grado de entrada, status, contrastatus y prestigio se han empleado en análisis de grafos dirigidos (Harary, 1965); las medidas de la distancia convertida y de centralidad han sido aplicadas por (Botafogo, 1992) y la textura ha sido definida por (Bernstein, 1992).

Tipo de Matriz Valor de la fila i, columna j	ADYACENCIA a_{ij}	DISTANCIA d_{ij}	DISTANCIA CONVERTIDA d'_{ij}
N1: Suma de los valores en la fila i	$\sum_{j=1}^{n} a_{ij} = d_i$ Grado de Apertura de v_i	$\sum_{j=1}^{n} d_{ij} = S_i$ Status de v_i	$\sum_{j=1}^{n} d'_{ij} = S'_i$ COD de v_i
N2: Suma de los valores en la columna j	$\sum_{i=1}^{p} a_{ij}$ Grado de entrada de v_j	$\sum_{i=1}^{p} d_{ij}$ Contrastatus de v_j	$\sum_{i=1}^{p} d'_{ij}$ CID de v_j
N3: Suma de los valores en la fila h menos la suma de los valores en la columna h	$\sum_{j=1}^{n} a_{hj} - \sum_{i=1}^{p} a_{ih}$	$\sum_{j=1}^{n} d_{hj} - \sum_{i=1}^{p} d_{ih}$ Prestigio (status de red) de v_h	$\sum_{j=1}^{n} d'_{hj} - \sum_{i=1}^{p} d'_{ih}$
N4: Ratio de la suma de valores en todas las filas por los valores en la fila h	$\frac{\sum_{i=1}^{p} d_i}{d_h}$	$\frac{\sum_{i=1}^{p} S_i}{S_h}$	$\frac{\sum_{i=1}^{p} S'_i}{S'_h}$ ROC de v_h

N5:		$\sum_{j=1}^{n} \frac{1}{2^{d_{ij}}}$ Textura de v_i (Bernstein, 1992)	
S_{Dice}	$\frac{2\sum\left(c_{jk} . c_{jl}\right)}{\sum\left(c_{jk}\right)^2 + \sum\left(c_{jl}\right)^2}$		
S_{cos}	$\frac{\sum\left(c_{jk} . c_{jl}\right)}{\sqrt{\sum\left(c_{jk}\right)^2 . \sum\left(c_{jl}\right)^2}}$		

3.3.2. Índices de grafo.

Con los índices de grafo obtenemos un valor para la similaridad del grafo completo, permitiendo la comparación entre grafos, siempre que exista un valor normalizado. Al igual que en el caso anterior, obtenemos un valor normalizado que refleja las características de un determinado grafo, permitiendo comparar el valor entre diferentes recogidas, pudiendo concluir que si el valor es el mismo ambos grafos son muy similares.

Tipo de Matriz Valor de la fila i, columna j	ADYACENCIA a_{ij}	DISTANCIA d_{ij}	DISTANCIA CONVERTIDA d'_{ij}
G1: Suma de valores en todas las filas (o en todas las columnas)	$\sum_{i=1}^{p} d_i$	$\sum_{i=1}^{p} s_i$	$\sum_{i=1}^{p} s'_i$ Distancia convertida de G
G2: Media de la suma de las filas	$\frac{\sum_{i=1}^{p} d_i}{p} = m$	$\frac{\sum_{i=1}^{p} s_i}{p}$	$\frac{\sum_{i=1}^{p} s'_i}{p}$

G3:			$\frac{p^3 - p^2 - \sum_{i=1}^{p} S'_i}{P^3 - 2P^2 - P}$ Compactación de G
G4:		$\sum_{h=1}^{p} \left\| \sum_{j=1}^{n} d_{hj} - \sum_{i=1}^{p} d_{ih} \right\|$ Suma de Prestigio absoluto de G	
G5: LAPS=$p^3/4$ si p es par LAPS=$(p^3-p)/4$ si p es impar		$\frac{\sum_{h=1}^{p} \left\| \sum_{j=1}^{n} d_{hj} - \sum_{i=1}^{p} d_{ih} \right\|}{LAPS}$ Stratum de G	
G6: la suma es sobre todos los nodos adyacentes	$\sum \left(d_i d_j \right)^{-1/2} = c$ Índice de conectividad de G (índice de Randic)		
G7:	$\frac{c}{r}$ Índice de Randic normalizado		

De todos los índice que hemos comentado nos vamos a centrar en los índice de grafo, que nos van a permitir comparar los diferentes dominios entre sí y el mismo dominio entre diferentes recogidas.

De estos índices de grafo nos vamos a centrar en la Compactación y el Stratum que nos van a permitir comparar la complejidad y conectividad de las estructuras hipertexto de los dominios Web (Botafogo, 1992). En contraposición

a la compactación analizamos el índice de Randic, como medida, que para grafos pequeños parece ofrecer los mismos resultados.

3.4. Medidas Topológicas de grafo.

3.4.1. Compactación.

El índice de Compactación de un grafo es una medida que se obtiene a partir de la matriz de distancia convertida y un grafo que posee un alto valor en su índice de compactación nos indica que los diferentes nodos pertenecientes al grafo se pueden alcanzar o enlazar fácilmente y ello sugiere un amplio número de referencias cruzadas o de enlaces entre los diferentes nodos. Un índice de compactación bajo indica que hay una insuficiencia de enlaces y que posiblemente existan diferentes partes del grafo que se encuentran desconectadas.

El índice de compactación se mueve entre el valor 0, indicando que la sede está totalmente desconectada, y el valor 1, indicando que la sede está totalmente conectada. Este índice es independiente del tamaño del grafo y siendo una medida extremadamente sensible a la más mínima variación en la estructura del grafo, que permite diferenciar grafos con el mismo número de nodos y de enlaces, pero en los que su estructura varía.

La compactación la podemos definir formalmente de la siguiente forma:

$$Cp = \frac{Max - \sum_i \sum_j C_{ij}}{Max - Min}$$

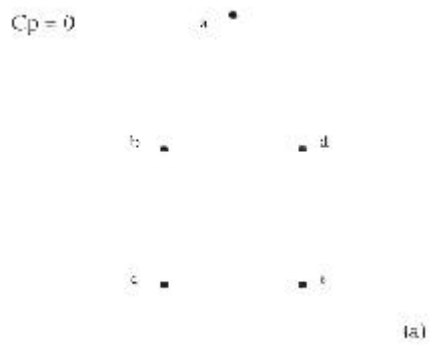
El cálculo de cada elemento es el siguiente:

$Max = (n^2 - n)C$; siendo n el número de nodos y C el máximo valor que puede tener un elemento de la matriz de distancia convertida que se corresponde precisamente con el número de nodos existentes.

$Min = (n^2 - n)$; siendo n el número de nodos.

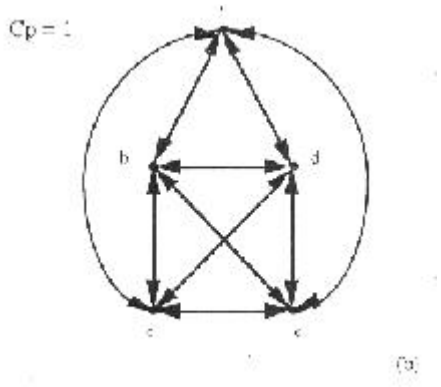
C_{ij} ; siendo la distancia convertida entre los nodos i y j.

A modo de ejemplo vamos a poner varios grafos y su correspondiente valor



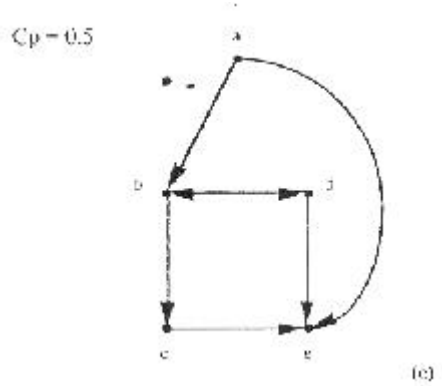
	a	b	c	d	e	COD
a	0	5	5	5	5	20
b	5	0	5	5	5	20
c	5	5	0	5	5	20
d	5	5	5	0	5	20
e	5	5	5	5	0	20
CID	20	20	20	20	20	100

Los valores para este grafo son: Max = 200; Min = 20; CD = 100.



	a	b	c	d	e	COD
a	0	1	1	1	1	4
b	1	0	1	1	1	4
c	1	1	0	1	1	4
d	1	1	1	0	1	4
e	1	1	1	1	0	4
CID	4	4	4	4	4	20

Los valores para este grafo son: Max = 100; Min = 20; CD = 20.



	a	b	c	d	e	COD
a	0	1	2	2	1	6
b	5	0	1	1	2	9
c	5	5	0	5	1	16
d	5	1	2	0	1	9
e	5	5	5	5	0	20
CID	20	12	10	13	5	60

Los valores para este grafo son: Max = 100; Min = 20; CD = 60.

Comenzando por la compactación, globalmente hay una ligera mejora en cuanto al aumento en la conexión de las diferentes sedes. En la primera recogida el 37% supera el 0,6 y en la segunda y tercera se mantiene en el 44,4%.

Analizando individualmente cada una de las sedes, en el gráfico adjunto (Figura 70), podemos ver un conjunto de sedes que en cada recogida mejoran su riqueza hipertextual, un 44,4% y en algún caso esta mejora es muy significativa como en el caso del *cicyt*, de la *upna*.

También hay que significar el importante descenso en la compactación de algunos dominios, como el caso de *ugr* o *uva*.

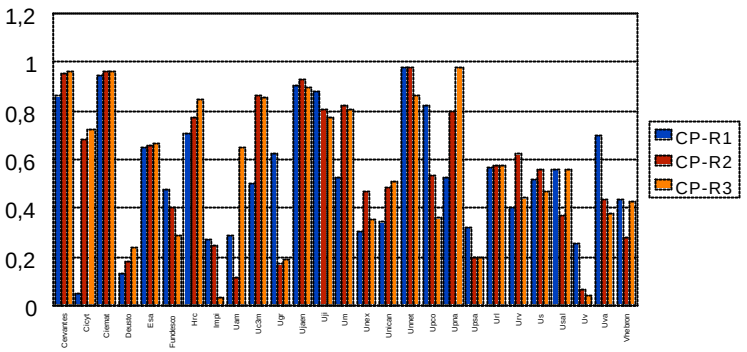


Figura 70: Valores de la Compactación en las tres recogidas.

Podemos concluir que las diferentes sedes, independientemente del crecimiento que puedan tener en cuanto a nodos y enlaces o la profundidad y distancia que tienen, están moderadamente conectadas, mejorando la estructura hipertextual entre las recogidas, pero con modificaciones drásticas en unos pocos casos, que han empeorado dicha estructura.

Los valores de cada sede, para cada recogida se muestran a continuación:

Dominio	CP-R1
Unnet	0,974924
Ciemat	0,944738
Ujaen	0,906451
Uji	0,878192
Cervantes	0,858994
Upco	0,820187
Hrc	0,708483
Uva	0,696973
Esa	0,652507
Ugr	0,626626
Url	0,567375
Usal	0,555965
Upna	0,528643
Um	0,528078
Us	0,519674

Dominio	CP-R2
Unnet	0,979432
Ciemat	0,958713
Cervantes	0,956726
Ujaen	0,925535
Uc3m	0,863493
Um	0,822218
Uji	0,809154
Upna	0,79819
Hrc	0,768814
Cicyt	0,681625
Esa	0,654034
Urv	0,620914
Url	0,578825
Us	0,559736
Upco	0,534897

Dominio	CP-R3
Upna	0,975327
Cervantes	0,961927
Ciemat	0,958754
Ujaen	0,895749
Unnet	0,866456
Uc3m	0,855803
Hrc	0,847964
Um	0,803155
Uji	0,77265
Cicyt	0,722129
Esa	0,662337
Uam	0,64845
Url	0,576113
Usal	0,557872
Unican	0,51064

Uc3m	0,503663
Fundesco	0,479902
Vhebron	0,436676
Urv	0,399558
Unican	0,341894
Upsa	0,320782
Unex	0,305776
Uam	0,290372
Impi	0,26962
Uv	0,258234
Deusto	0,136415
Cicyt	0,048445

Unican	0,482356
Unex	0,471501
Uva	0,434365
Fundesco	0,404423
Usal	0,372803
Vhebron	0,277657
Impi	0,246566
Upsa	0,198654
Deusto	0,183459
Ugr	0,169506
Uam	0,117727
Uv	0,066093

Us	0,465444
Urv	0,440266
Vhebron	0,428605
Uva	0,375291
Upco	0,358244
Unex	0,353867
Fundesco	0,285371
Deusto	0,238029
Upsa	0,198757
Ugr	0,186601
Uv	0,04299
Impi	0,031973

3.4.2. Índice de Randic.

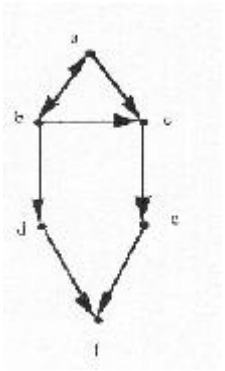
El índice de Randic es un índice que se comenzó a utilizar en química para caracterizar diferentes modelos de moléculas (Randic, 1975). En principio y para grafos de pocos nodos, indica exactamente lo mismo que el índice de compactación, pero en este caso los valores son justamente al contrario, tendiendo a 0 cuando está totalmente conectado y a 1 cuando está desconectado.

La comparación de estas dos medidas es importante, porque la obtención del índice de Randic, que se obtiene a partir de la matriz de adyacencia, es mucho menos costosa en tiempo de procesamiento que la compactación, que se obtiene a partir de la matriz de distancia.

Su cálculo es $\sum \left(d_i d_j\right)^{-1/2} = c$, donde d_i es el grado de apertura del vértice v_i y la suma es para todos los vértices adyacentes v_i y v_j. El índice de Randic así obtenido nos ofrece un valor no normalizado que varía en relación al número de aristas del grafo.

Por ello es necesario normalizar este valor por el número de enlaces del grafo, obteniendo valores normalizados que pueden servir para comparar datos de

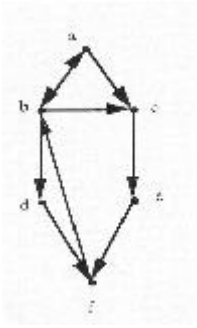
diferentes dominios o de diferentes recogidas. El índice de Randic quedaría formulado de la siguiente forma $\frac{c}{r}$, siendo r el número de enlaces del grafo.



Vamos a realizar una comparación del rendimiento del índice de compactación y del índice de Randic, analizando la variación en sus valores con la modificación de un único enlace.

	a	b	c	d	e	f
a	0	1	1	0	0	0
b	1	0	1	1	0	0
c	0	0	0	0	1	0
d	0	0	0	0	0	1
e	0	0	0	0	0	1
f	0	0	0	0	0	0

Para este grafo los valores son: Compactación = 0,42; Randic = 0,648



En este segundo grafo añadimos un nuevo enlace entre f y b dando la siguiente matriz y resultados.

	a	b	c	d	e	f
a	0	1	1	0	0	0
b	1	0	1	1	0	0
c	0	0	0	0	1	0
d	0	0	0	0	0	1
e	0	0	0	0	0	1
f	0	1	0	0	0	0

Para este grafo los valores son: Compactación = 0,7667; Randic = 0,6316

Como podemos ver, aunque son dos medidas que teóricamente miden lo mismo, en el caso del índice de compactación los valores obtenidos son mucho más sensibles, respecto al índice de Randic. En grafos de muy pocos nodos las modificaciones en un solo enlace se reflejan automáticamente tanto en el índice de compactación como en el de Randic, pero en grafos de un número elevado de nodos el índice de compactación refleja inmediatamente cualquier modificación mientras el índice de Randic lo refleja mucho peor. Posiblemente al trabajar el índice de compactación con las distancias, la inclusión o la modificación de un solo enlace modifica inmediatamente las distancias calculadas hasta ese momento y quedando reflejadas por dicho índice.

El algoritmo para calcular este índice sería el siguiente:

```
for j = 1:f          **siendo f el número de nodos
        for k = 1:c          **siendo c el número de nodos
                if Y(j,k) = = 1 **siendo Y la matriz de adyacencia
                        Ran = (DO(j)*DI(k)) ^ (-0.5); **siendo DO el grado de
                                                        apertura y DI el grado de
entrada
                        Randic = Randic + Ran;
                end
        end
end
RandicNor = Randic/P **siendo P el número de enlaces
```

El índice de Randic, o de conectividad, sigue el proceso inverso al del índice de compactación, tiende a cero las sedes más conectadas. El 48% de las sedes mejora su conectividad en todas las recogidas, estando el 26% de ellas muy bien conectadas en la última recogida. Se aprecia igualmente una pérdida de rendimiento en las sedes ugr, uva.

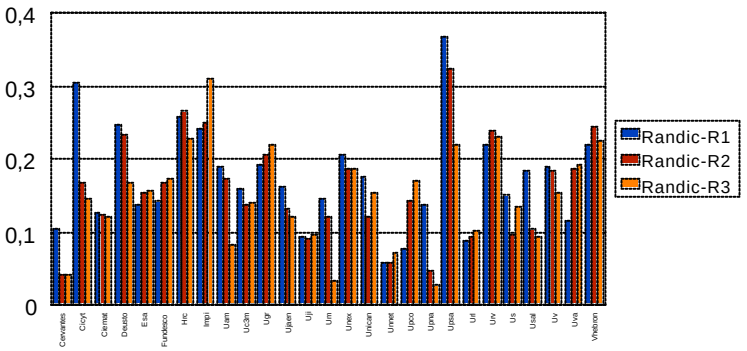


Figura 71: Valores del Randic normalizado en las tres recogidas.

Los valores de cada recogida son los siguientes:

Dominio	Randic-R1
Unnet	0,058699
Upco	0,077174
Url	0,087783
Uji	0,093038
Cervantes	0,105966
Uva	0,114981
Ciemat	0,125607
Upna	0,137744
Esa	0,137962
Fundesco	0,141698
Um	0,144903
Us	0,149909
Uc3m	0,159994
Ujaen	0,163036
Unican	0,174894
Usal	0,183463
Uv	0,189191
Uam	0,190137
Ugr	0,1927
Unex	0,207034
Urv	0,219745
Vhebron	0,220476
Impi	0,240839
Deusto	0,246157
Hrc	0,256944
Cicyt	0,304423
Upsa	0,36709

Dominio	Randic-R2
Cervantes	0,042689
Upna	0,046518
Unnet	0,057711
Uji	0,091522
Url	0,093789
Us	0,097742
Usal	0,103479
Unican	0,122133
Um	0,122326
Ciemat	0,122814
Ujaen	0,131226
Uc3m	0,136947
Upco	0,142153
Esa	0,154978
Fundesco	0,166668
Cicyt	0,168948
Uam	0,173088
Uv	0,182775
Unex	0,186099
Uva	0,187214
Ugr	0,206411
Deusto	0,232904
Urv	0,238664
Vhebron	0,244234
Impi	0,248247
Hrc	0,264537
Upsa	0,322608

Dominio	Randic-R3
Upna	0,026966
Um	0,034049
Cervantes	0,040886
Unnet	0,07276
Uam	0,082863
Usal	0,09342
Uji	0,097826
Url	0,1023
Ujaen	0,120689
Ciemat	0,121687
Us	0,134569
Uc3m	0,139106
Cicyt	0,145341
Unican	0,153123
Uv	0,154076
Esa	0,155915
Deusto	0,167599
Upco	0,169283
Fundesco	0,172683
Unex	0,186095
Uva	0,191151
Ugr	0,218558
Upsa	0,220479
Vhebron	0,224816
Hrc	0,228483
Urv	0,231214
Impi	0,310564

3.4.3. Stratum.

Stratum, es un índice que nos permite conocer si el hipertexto se ha diseñado de una forma lineal, jerárquica, induciendo al usuario a seguir un orden concreto en la obtención de la información o si por el contrario no existe esta estructura jerárquica y no existe un orden preestablecido por el diseñador de la sede para poder obtener la información.

Los valores que puede alcanzar el stratum van desde el valor 1, indicando una estructura hipertextual de tipo lineal, que no es interesante al no aprovechar las ventajas del hipertexto; hasta el valor 0, que nos indica una estructura circular en la que no se diferencia el nodo por el que poder comenzar la lectura de la información.

El cálculo del stratum se basa en los conceptos de Status y ContraStatus, definidos por (Harary, 1965) y que pueden calcularse de la siguiente forma:

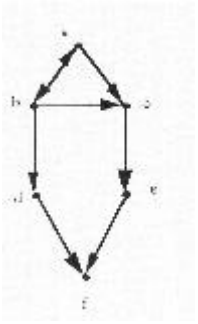
El Status es $\sum_{j=1}^{n} d_{ij} = S_i$, es decir la suma de todos los elementos de una fila en la matriz de distancia y el ContraStatus es $\sum_{i=1}^{p} d_{ij}$, es decir la suma de todos los elementos de una columna en la matriz de distancia. Harary indicaba, que una medida complementaria podía ser el Prestigio, denominado Status de Red por (Ellis, 1994), calculándolo de la siguiente forma $\sum_{j=1}^{n} d_{hj} - \sum_{i=1}^{p} d_{ih}$.

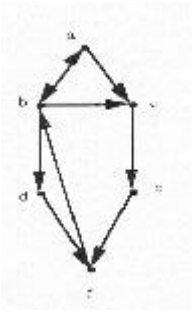
El Prestigio total de un grafo siempre es cero, al ser los valores de Status total y Contrastatus total idénticos. Por ello se ha definido el Prestigio absoluto, realizando para ello la suma de los valores absolutos del Status.

El Prestigio Absoluto Lineal (LAP) de un hipertexto con n nodos es idéntico al prestigio absoluto de un hipertexto lineal de n nodos y se calcula de la siguiente forma $LAP = \begin{cases} \frac{n^3}{4} \, si\ el\ número\ de\ nodos\ es\ par \\ \frac{n^3 - n}{4} \, si\ el\ número\ de\ nodos\ es\ impar \end{cases}$

Con estos datos podemos definir el Stratum como $St = \frac{prestigio\ absoluto}{LAP}$ consiguiendo una medida normalizada independiente del número de nodos.

Para valorar la potencia de esta medida, vamos a indicar los cálculos para los dos grafos siguientes:





	a	b	c	d	e	f	Stat.	Prest.
a	0	1	1	2	2	3	9	8
b	1	0	1	1	2	2	7	6
c	0	0	0	0	1	2	3	1
d	0	0	0	0	0	1	1	-2
e	0	0	0	0	0	1	1	-4
f	0	0	0	0	0	0	0	-9
CStat.	1	1	2	3	5	9		

Prestigio Absoluto 30

El Stratum para este grafo con un LAP de 54 ($6^3/4$) es St=0,56

	a	b	c	d	e	f	Stat.	Prest.
a	0	1	1	2	2	3	9	-4
b	1	0	1	1	2	2	7	-2
c	4	3	0	4	1	2	14	4
d	3	2	3	0	4	1	13	1
e	3	2	3	3	0	1	12	0
f	2	1	2	2	3	0	10	1
CStat.	1	1	2	3	5	9		

Prestigio Absoluto 12

El Stratum para este grafo con un LAP de 54 ($6^3/4$) es St = 0,22.

Como podemos ver, ante dos grafos prácticamente idénticos con la excepción de un único enlace, el stratum se modifica sustancialmente, dándonos una idea de su precisión.

En el análisis de los datos obtenidos, podemos observar que el valor máximo que se tiene se encuentra por debajo de 0,1 que nos indica que todas las sedes en conjunto carecen de una estructura jerárquica y de una premeditación por parte del diseñador de la sede en inducir el orden en el que el usuario debe obtener la información.

Exceptuando unas pocas sedes, todas mantienen un nivel de stratum por debajo de 0,02 con muy pocas variaciones entre recogidas y en general manteniendo valores muy similares.

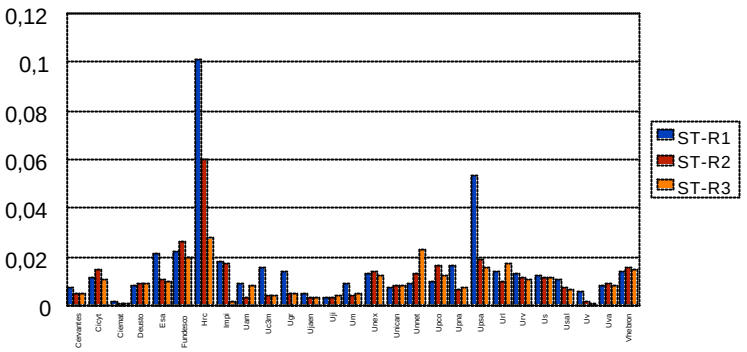


Figura 72: Valores del Stratum en las tres recogidas.

Dominio	ST-R1
Hrc	0,101449
Upsa	0,053339
Fundesco	0,022173
Esa	0,021407
Impi	0,018508
Upna	0,016522
Uc3m	0,01577
Vhebron	0,014381
Url	0,013835
Ugr	0,013827
Unex	0,013238
Urv	0,013158
Us	0,012744
Cicyt	0,012026
Usal	0,010627
Upco	0,010129
Unnet	0,009403
Uam	0,009353
Um	0,009257
Deusto	0,00849
Uva	0,008196

Dominio	ST-R2
Hrc	0,060169
Fundesco	0,0264
Upsa	0,019074
Impi	0,017689
Upco	0,016724
Vhebron	0,015885
Cicyt	0,015045
Unex	0,014157
Unnet	0,013032
Urv	0,012015
Us	0,0117
Esa	0,011044
Url	0,010119
Uva	0,009574
Deusto	0,009135
Unican	0,008528
Usal	0,007896
Upna	0,007165
Ugr	0,004901
Cervantes	0,004874
Um	0,004472

Dominio	ST-R3
Hrc	0,027677
Unnet	0,022888
Fundesco	0,019597
Url	0,017379
Upsa	0,016117
Vhebron	0,01499
Upco	0,012612
Unex	0,012489
Us	0,01141
Urv	0,01127
Cicyt	0,011084
Esa	0,010314
Deusto	0,009256
Unican	0,008716
Uva	0,008589
Uam	0,008292
Upna	0,007545
Usal	0,006777
Ugr	0,005245
Um	0,004971
Cervantes	0,004853

Unican	0,007616
Cervantes	0,007257
Uv	0,005852
Ujaen	0,005366
Uji	0,003327
Ciemat	0,001944

Uc3m	0,004207
Uji	0,003788
Ujaen	0,003547
Uam	0,003485
Uv	0,00164
Ciemat	0,001409

Uc3m	0,004625
Uji	0,004237
Ujaen	0,003538
Impi	0,001803
Ciemat	0,001304
Uv	0,001012

3.5 Conclusiones.

El manejo del Web como si fuese un grafo, es aceptado mayoritariamente como uno de los mejores mecanismos existentes para estudiar y analizar el Web.

La cantidad de estudios de todo tipo que aplican estas ideas es enorme y un pequeño reflejo de los mismos se ha ofrecido a lo largo de este capítulo.

La variedad de índices que se pueden aplicar es realmente muy alta, como ya indicamos y nosotros nos centramos en unos pocos de los índices que afectan al análisis de l grafo completo y no a los que afectan a los nodos.

Comenzando con la compactación, que es un índice que nos indica el nivel de conectividad existente entre los diferentes nodos del grafo, podemos concluir que de forma global los grafos que conforman cada dominio mejoran su conectividad. De forma individual hay una serie de dominios que mejoran constantemente su valor de compactación, siendo en algún caso muy significativa, como *cicyt* o *upna* y de la misma forma, algunos dominios como *ugr* o *uva* empeoran sus resultados de forma drástica.

El índice de Randic, que en teoría sirve para medir lo mismo que el índice de compactación, muestra la misma tendencia observada en el índice de compactación. Sin embargo, así como en grafos muy pequeños, la modificación en un solo enlace se refleja igualmente en ambos índices, en grafos grandes, el índice de Randic es menos preciso en el reflejo de dicho cambio, frente al índice de compactación, en el que se refleja dicho cambio con mucha precisión.

Evidentemente, al calcular ambos índices a partir de diferente tipo de matriz, tiene su reflejo en el valor obtenido. El índice de compactación al basarse en la

matriz de distancia, es muy sensible a cualquier cambio en los enlaces, pues lleva implícita la modificación de la distancia a la que se encuantran los diferentes nodos.

En relación a stratum, que nos permite conocer si el diseño del hipertexto es jerárquico o no. El diseño jerárquico implica que el diseñador del hipertexto induce al usuario a obtener la información de una determinada forma.

En este sentido todos los datos nos indican que todos los dominios tienen una estructura hipertexto no jerárquica y no existen recorridos preestablecidos por el diseñador del hipertexto para poder obtener la información. Los valores de cada dominio en las diferentes recogidas se mantienen bastante similares.

4. Leyes de Exponenciación.

4.1. Introducción.

Aparentemente el Web crece de forma aleatoria y sin mecanismos que de alguna forma regulen dicho crecimiento. Sin embargo, se han descubierto unas leyes muy sencillas que indican que la topología Web sigue algunas pautas de funcionamiento que son interesantes para analizar el Web y que pueden ser utilizadas para su análisis. (Huberman, 1999) apunta la existencia de unas leyes que gobiernan el crecimiento dinámico del Web, y la denomina ley de exponente, pero sin entrar en mayores consideraciones.

Los estudios de (Faloutsos, 1999) han determinado que las topologías Web siguen leyes del tipo $y = x^a$, (similares a la de la ley de Zipf (Zipf, 1949)) dando lugar a cuatro leyes, que caracterizan dicha topología. Este fenómeno ha sido observado también en los trabajos de (Huberman, 1999), (Kleinberg, 1999b), (Kumar, 1999), y (Kumar, 1999b), aunque la definición de dichas leyes y su soporte teórico pertenece a (Faloutsos, 1999)

Cada una de las leyes se caracteriza por tener un valor único para todos los datos analizados, y este valor es un exponente, que es el valor de la pendiente que tiene la representación gráfica de los datos que se analizan en cada una de las leyes y que nos va a permitir identificar diferentes grafos (Faloutsos, 1999) al ser valores normalizados. Este valor único que nos ofrece cada una de las leyes nos permite comparar con los valores obtenidos para otros grafos o dominios y saber si son muy parecidos o no.

Símbolo	Definición
G	Grafo
N	Número de nodos en el grafo
E	Número de aristas en el grafo
d	Diámetro del grafo

d_v	Grado de apertura del nodo v, definido como $\sum_{j=1}^{n} a_{ij} = d_i$, es decir la suma de todos los valores de una fila que nos indica lo bien o mal conectado que se encuentra un determinado nodo
$\overline{d}$	Media aritmética del grado de apertura de los nodos del grado, definida como $\overline{d} = 2E / N$
f_d	Frecuencia de un grado de apertura d, que es el número de nodos con el grado de apertura d
r_v	Orden del nodo v, que es un índice en orden decreciente del grado de apertura
P(h)	Número de pares de nodos con menor o igual número de saltos
l	Valores propios de la matriz
i	El orden de l_i

Vamos a estudiar cada una de las leyes desde el punto de vista teórico y analizaremos los resultados de cada una de las leyes propuestas por (Faloutsos, 1999) indicando las peculiaridades observadas en los dominios españoles analizados.

Estas leyes empiezan a ser ampliamente estudiadas y uno de los mejores trabajos es el de (Medina, 2000) que mediante generación de topologías de red en laboratorio ha extraído algunas conclusiones interesantes de las mismas. Estas conclusiones se basan en dos características observadas por Medina que a continuación exponemos desde el punto de vista teórico:

1. Crecimiento exponencial en la topología de red. La presencia de esta característica nos permite disponer de redes abiertas que aceptan nuevos nodos continuamente.

2. Conectividad preferencial. La presencia de esta característica indica una tendencia de los nuevos nodos a conectarse a nodos existentes que tienen un alto grado de apertura.

El trabajo de (Palmer, 2000) incide en las aportaciones de (Medina, 2000) y ofrece como novedoso algunos algoritmos para el cálculo de algunas de las leyes.

(Shiode, 2000) presenta un trabajo de aplicación de alguna de las leyes de exponenciación y confirmando en sus resultados su existencia.

Todas las leyes, en función de los datos que se recogen y se manejan para cada una de ellas, se representan gráficamente y obtenemos dos valores que se consideran importantes:

1. El valor de la pendiente de la gráfica y ese valor será el valor del exponente analizado.

2. El Coeficiente de correlación de los datos. Para manejarlo se utiliza el coeficiente de correlación absoluto (CCA), que nos va a indicar si los datos que estamos manejando presentan o no la ley que estamos tratando de comprobar. Según (Faloutsos, 1999) los dominios que disponen de un CCA superior o igual a 0,95 presentan la ley que estamos analizando. Por esto, en los resultados obtenidos de nuestra investigación manejamos este valor como umbral.

4.2. Exponente de orden R (Ley de exponente 1).

En esta primera ley se estudian los grados de apertura de los diferentes nodos que forman parte del grafo, que representa a los dominios Web. Este grado de apertura, como ya hemos indicado nos da una idea de lo bien o mal conectado que se encuentra un determinado nodo.

Los nodos que poseen un grado de apertura mayor están más relacionados, poseen más enlaces que aquellos que tienen un valor menor. Para poder analizar esta ley, ordenaremos los grados de apertura de los nodos objeto de estudio en orden decreciente y realizaremos la gráfica correspondiente, que nos ofrecerá una línea de tendencia con una pendiente, que será precisamente el valor de este exponente (Faloutsos, 1999).

El valor de esta ley nos indica, comparándolo con otros valores, si los diferentes dominios analizados son muy parecidos o no, desde el punto de vista de la conectividad de los nodos (Pansiot, 1998).

Podemos definir formalmente la Ley 1 de la siguiente forma (Faloutsos, 1999): el grado de apertura, d_v, de un nodo v, es proporcional al orden del nodo, r_v, elevado a un exponente, R,

$$d_v \; a \; r_v^R$$

El exponente de orden R, es la pendiente que se obtiene con la representación del grado de apertura de los nodos frente al orden de los nodos en una escala logarítmica.

Una de las utilidades de este exponente, es que nos permite comparar diferentes topologías, diferenciando diferentes representaciones del grafo Web, atendiendo a la conectividad de los diferentes nodos.

De esta primera ley se pueden sacar algunos lemas que completan dicha ley y nos ofrecen nuevos valores útiles en la caracterización de dicha topología.

En primer lugar, si consideramos que el mínimo grado de apertura de un nodo es 1 ($d_N = 1$), podemos decir que el grado de apertura, d_v, de un nodo v, es una función del orden del nodo, r_v, y del exponente R de la siguiente forma $d_v = \frac{1}{N^R} r_v^R$.

Aplicando este lema podemos relacionar el número de aristas con el número de nodos, N, y el exponente R

$$E = \frac{1}{2\,(R+1)} \left(1 - \frac{1}{N^{R+1}} \right) N$$

Se ha estimado que el número de aristas obtenidas mediante la aplicación de este lema, difiere entre 9-20% de los datos reales (Faloutsos, 1999).

La ley de exponente 1 necesita para aparecer, según (Medina, 2000) de un crecimiento exponencial en la topología de red y de una conectividad preferencial y por ello los dominios que no tengan estas características no cumplirán con la Ley 1.

Los datos obtenidos en la investigación son los siguientes.

4.2.1. Primera recogida.

El 33% de los dominios presenta un coeficiente de correlación absoluto (CCA) > = 0,95 indicando la presencia de dicha ley en su topología internet y por ello se ajustan a los valores obtenidos por (Faloutsos, 1999) y dichos dominios ofrecen un crecimiento exponencial y una conectividad preferencial muy acusada (Medina, 2000), ofreciendo unos buenos niveles de conectividad.

El 40,7% de los dominios tiene un CCA > = 0,9 y < = 0,95 indicándonos que algunas de las características determinantes en la presencia de esta ley tienen un menor peso en su implantación.

El 26,3% de los dominios no presenta esta ley y por ello carecen de las características mencionadas anteriormente.

Respecto al valor de la pendiente obtenida por los dominios con CCA más alto, se mantiene por encima de los valores obtenidos en (Faloutsos, 1999) pero manteniendo una cierta disparidad en sus valores, indicándones que la distribución del exponente *R* en los dominios españoles es bastante dispar y la evolución de las diferentes topologías no se realiza atendiendo a ningún mecanismo común en su crecimiento y desarrollo.

Dominio	Correlacion-R1	Pendiente-R1
Impi	-0,970244	-0,724856
Uv	-0,965725	-0,980656
Ujaen	-0,965578	-0,6064
Urv	-0,954314	-0,908705
Ugr	-0,953725	-0,782257
Unex	-0,95271	-0,93105
Vhebron	-0,952193	-0,823591
Uc3m	-0,948364	-0,801423
Cicyt	-0,947974	-1,71737

Dominio	Correlacion-R1	Pendiente-R1
Hrc	-0,944153	-0,541507
Esa	-0,942108	-0,698906
Uam	-0,938732	-0,947454
Uji	-0,936537	-0,531549
Us	-0,934658	-0,984662
Unican	-0,924229	-0.925426
Um	-0,922119	-0,870108
Usal	-0,910609	-0,872829
Upna	-0,90692	-0,808997
Deusto	-0,901611	-0,808032
Upsa	-0,895959	-0,641732
Uva	-0,884216	-0,766409
Ciemat	-0,87173	-0,431489
Cervantes	-0,86298	-0,684682
Fundesco	-0,862449	-0,795136
Upco	-0,817767	-0,70421
Url	-0,800153	-0,865244
Unnet	-0,766682	-0,675043

Los gráficos para el mejor y el peor dominio son los siguientes:

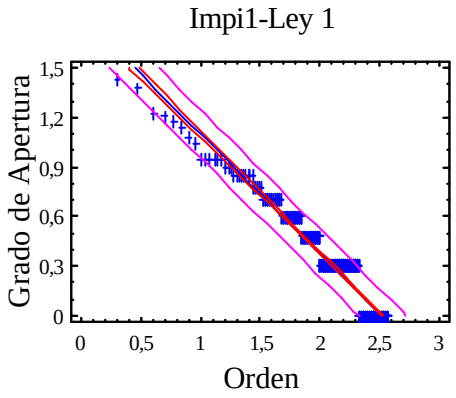


Figura 73: Mejor CCA en Ley 1-R1

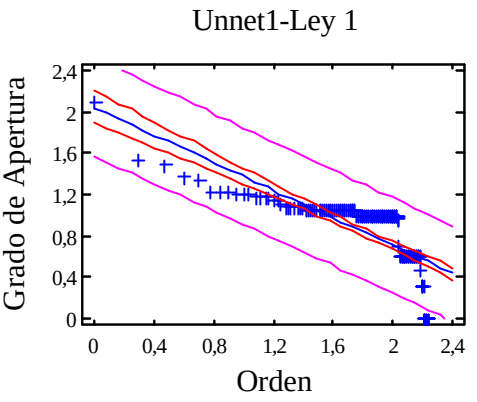


Figura 74: Peor CCA en Ley 1-R1

4.2.2. Segunda recogida.

El 26% de los dominios presenta un coeficiente de correlación absoluto (CCA) >= 0,95 indicando una pérdida en el crecimiento exponencial y en la conectividad preferencial y por ello obteniendo unas topologías peores, desde el punto de vista de la conectividad de los nodos.

El 33% de los dominios tiene un CCA >=0,9 y <= 0,95 siguiendo la misma línea apuntada en los datos anteriores.

El 41% de los dominios no presenta esta ley.

Podemos observar una perdida de eficacia en la conectividad de los nodos implicando que el diseño de las topologías Internet en esta segunda recogida es peor respecto a la primera.

Respecto al valor de la pendiente obtenida por los dominios con CCA más alto, ahora los valores son ligeramente más bajos que los obtenidos por (Faloutsos, 1999) pero los valores se mantienen más estables y sin tanta disparidad, indicando un buen nivel de correlación y manteniendo unas topologías más similares que en la primera recogida.

Dominio	Correlacion-R2	Pendiente-R2
Impi	-0,974701	-0,761987

Dominio	Correlacion-R2	Pendiente-R2
Uc3m	-0,97005	-0,730643
Hrc	-0,964329	-0,778
Uv	-0,958582	-1,0129
Urv	-0,95745	-0,805183
Esa	-0,9499	-0,742463
Um	-0,947177	-0,655643
Ujaen	-0,939478	-0,522932
Ugr	-0,939211	-0,892451
Vhebron	-0,93617	-0,825306
Unex	-0,935256	-0,874928
Upsa	-0,934551	-0,795522
Uva	-0,921557	-0,905971
Uam	-0,91393	-0,94063
Deusto	-0,898873	-0,778569
Ciemat	-0,896758	-0,412428
Unican	-0,890321	-0,780314
Fundesco	-0,878562	-0,789272
Upco	-0,871443	-0,671023
Uji	-0,856194	-0,496968
Us	-0,835433	-0,935923
Usal	-0,83133	-0,745027
Url	-0,807278	-0,921528
Cicyt	-0,790076	-0,493027
Unnet	-0,75476	-0,647836
Upna	-0,686583	-0,602229
Cervantes	-0,681058	-0,456687

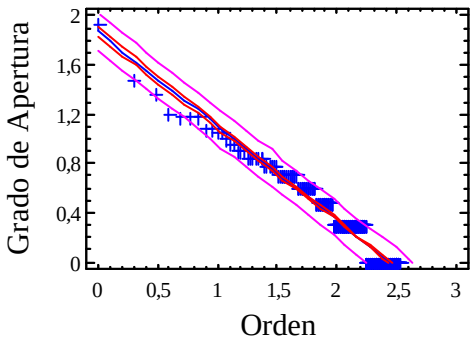


Figura 75: Mejor CCA en Ley 1-R2

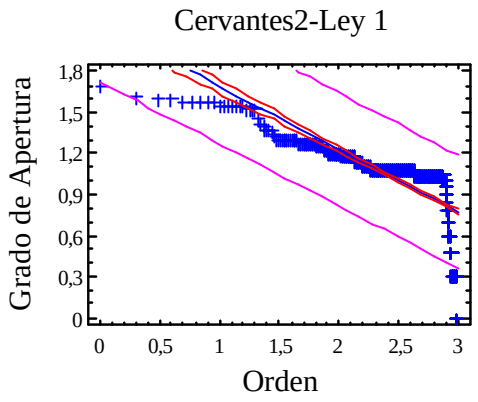


Figura 76: Peor CCA en Ley 1-R2

4.2.3. Tercera recogida.

El 26% de los dominios presenta un coeficiente de correlación absoluto (CCA) >= 0,95 indicando un mantenimiento de los valores obtenidos en la recogida anterior, pero además se trata de los mismos dominios que en la segunda recogida (salvo um) y aparece ugr (que mantiene unos niveles altos en todas las recogidas). Lo importante es que todos estos dominios mejoran su CCA, indicando que las topologías se han asentado y tienden a mejorar en sus valores.

El 30% de los dominios tiene un CCA >=0,9 y <= 0,95 perdiendo ligeramente porcentaje, pero en este caso, los valores en los dominios coincidentes han supuesto una pérdida respecto a la segunda recogida.

El 44% de los dominios no presenta esta ley suponiendo una pérdida respecto a las recogidas anteriores.

Podemos observar una perdida de eficacia en el diseño de las topologías Internet en esta tercera recogida respecto a la segunda, pero los dominios de mejores valores mejoran sus resultados..

Respecto al valor de la pendiente obtenida por los dominios con CCA más alto, los valores vuelven a ser un poco más dispares.

Realizando una valoración global, podemos observar que los dominios que en la primera recogida tenían un CCA > = 0,95 la mayoría mantienen ese status en todas las recogidas (excepto alguno de los dominios) y en general su CCA mejora al final de las mismas. Algunos de los nuevos dominios que se englobarían dentro de ese grupo en las recogidas posteriores provienen de dominios que en la primera recogida se engloban dentro del grupo de CCA > = 0,9 y < = 0,95 y que por lo tanto han modificado a mejor su topología de red.

Dominio	Correlacion-R3	Pendiente-R3
Uv	-0,981441	-1,11601
Impi	-0,981185	-1,03769
Uc3m	-0,972502	-0,70861
Ugr	-0,966435	-0,887819
Esa	-0,962321	-0,766517
Hrc	-0,959702	-0,787982
Urv	-0,952787	-0,909852
Ujaen	-0,931355	-0,598113
Unex	-0,924851	-0,889572
Vhebron	-0,923067	-0,845697
Uva	-0,915773	-0,893299
Upsa	-0,906101	-0,843371
Deusto	-0,902617	-0,946887
Us	-0,902555	-1,02438
Ciemat	-0,901309	-0,426752
Unican	-0,890481	-0,792638
Upco	-0,882774	-0,681378
Uji	-0,878107	-0,520192
Fundesco	-0,866401	-0,772396
Upna	-0,853166	-0,43079
Cicyt	-0,844663	-0,616466
Usal	-0,836916	-0,78967
Url	-0,79707	-0,857801
Uam	-0,788171	-0,502392
Unnet	-0,786417	-0,735048
Cervantes	-0,699771	-0,457316
Um	-0,582266	-0,553687

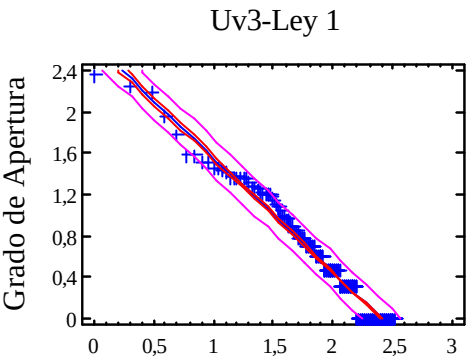


Figura 77: Mejor CCA en Ley 1-R3

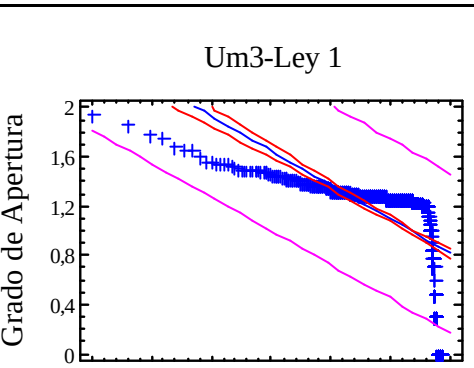


Figura 78: Peor CCA en Ley 1-R3

4.3. Exponente de grado de apertura O (Ley de exponente 2).

Para el análisis de esta ley manejamos también el grado de apertura, pero en este caso los diferentes grados de apertura se agrupan y se calcula la frecuencia de cada uno de ellos. Analizamos la distribución del grado de apertura de los grafos analizados. La presencia de esta ley será una indicación de que la distribución de

los grados de apertura de los nodos no es arbitraria y que sigue algún tipo de patrón (Faloutsos, 1999).

Podemos definir formalmente la Ley 2 de la siguiente forma: la frecuencia, f_d, con un grado de apertura, d, es proporcional al grado de apertura elevado a un exponente, O.

$$f_d \; a \; d^{O}$$

El exponente del grado de apertura O, es la pendiente que se obtiene con la representación de la frecuencia del grado de apertura frente a los grados de apertura en una escala logarítmica.

La presencia de esta ley indica que la distribución de los grados de apertura de los nodos Web no es arbitraria, y los nodos con un grado de apertura bajo son más frecuentes.

Los trabajos de (Medina, 2000) indican que para la ley de exponente 2 también se necesita que exista un crecimiento exponencial en la topología de red, permitiendo redes abiertas que aceptan nuevos nodos continuamente y de una conectividad preferencial, indicando una tendencia de los nuevos nodos a conectarse a nodos existentes con un alto grado de apertura.

Los datos obtenidos en la investigación son los siguientes:

4.3.1. Primera recogida.

Ninguno de los dominios presenta un coeficiente de correlación absoluto (CCA) > = 0,95 indicando que esta ley no queda bien representada con este exponente en los dominios españoles. Si tenemos en cuenta que el porcentaje de implantación de la Ley 1 no es demasiado fuerte, al representar esta ley mediante la frecuencia del grado de apertura, nos indica que los dominios españoles no siguen ninguna pauta en cuanto a esa frecuencia del grado de apertura.

El 30% de los dominios tiene un CCA > = 0,9 y < = 0,95 indicándonos que algunas de las características determinantes en la presencia de esta ley tienen un menor peso en su implantación.

El 70% de los dominios no presenta esta ley y por ello carecen de las características mencionadas anteriormente.

Respecto al valor de la pendiente obtenida por los dominios, está muy por debajo de los valores obtenidos en (Faloutsos, 1999) pero manteniéndose en valores bastante similares, indicándonos que la distribución del exponente O en los dominios españoles es bastante similar.

Dominio	Correlacion-R1	Pendiente-R1
Us	-0,92805	-1,38849
Vhebron	-0,919401	-1,66087
Uam	-0,917655	-1,35978
Unex	-0,913727	-1,27117
Urv	-0,911492	-1,43218
Unican	-0,910294	-1,47402
Ugr	-0,901895	-1,45181
Impi	-0,898572	-1,5168
Uva	-0,894394	-1,68559
Upna	-0,892806	-1,45638
Usal	-0,890889	-1,51481
Uc3m	-0,890528	-1,57427
Ujaen	-0,889016	-1,54251
Um	-0,871307	-1,38783
Uv	-0,869914	-1,07842
Esa	-0,85019	-1,16626
Upco	-0,847865	-1,26876
Fundesco	-0,842618	-1,36317
Upsa	-0,797175	-1,16586
Deusto	-0,795502	-1,20279
Uji	-0,777111	-1,3138
Hrc	-0,760285	-1,57409
Cervantes	-0,751608	-1,40016
Url	-0,740101	-1,13464
Ciemat	-0,726103	-1,08657
Cicyt	-0,659483	-0,157204
Unnet	-0,44779	-0,559355

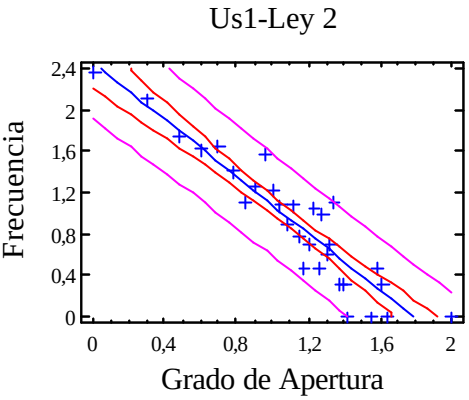


Figura 79: Mejor CCA en Ley 2-R1

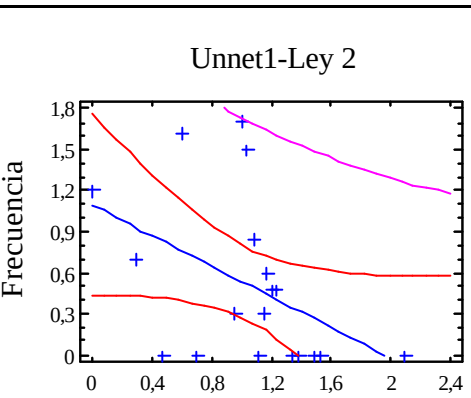


Figura 80: Peor CCA en Ley 2-R1

4.3.2. Segunda recogida.

Ninguno de los dominios presenta un coeficiente de correlación absoluto (CCA) > = 0,95 indicando que esta ley no queda bien representada con este exponente en los dominios españoles.

El 40,7% de los dominios tiene un CCA >=0,9 y <= 0,95 indicándonos que algunas de las características determinantes en la presencia de esta ley tienen un menor peso en su implantación.

El 59,3% de los dominios no presenta esta ley y por ello carecen de las características mencionadas anteriormente.

Respecto al valor de la pendiente obtenida por los dominios, está muy por debajo de los valores obtenidos en (Faloutsos, 1999) pero manteniéndose en valores bastante similares, indicándonos que la distribución del exponente *O* en los dominios españoles es bastante similar.

Dominio	Correlacion-R2	Pendiente-R2
Hrc	-0,94389	-1,33845
Uva	-0,937587	-1,64235
Esa	-0,932595	-1,56883
Urv	-0,923842	-1,54201
Vhebron	-0,923643	-1,677
Unex	-0,918959	-1,67737
Upsa	-0,912896	-1,27748
Ugr	-0,912266	-1,512226
Impi	-0,90253	-1,39291
Uam	-0,898912	-1,30701
Uc3m	-0,898679	-1,452203
Unican	-0,884994	-1,46876
Upco	-0,877774	-1,73303
Deusto	-0,868757	-1.5816
Fundesco	-0,867506	-1,54036
Um	-0,862121	-1,32874
Us	-0,848566	-1,23805
Ujaen	-0,825961	-0,452906
Uv	-0,822894	-0,886963
Usal	-0,811094	-1,29205
Uji	-0,798982	-1,36519
Cicyt	-0,788312	-1,35654
Ciemat	-0,716257	-1,03716
Url	-0,710218	-1,14699
Upna	-0,701103	-1,02339
Cervantes	-0,52342	-0,920653
Unnet	-0,425361	-0,522755

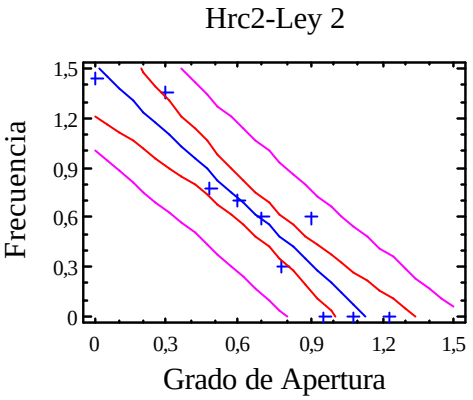


Figura 81: Mejor CCA en Ley 2-R2

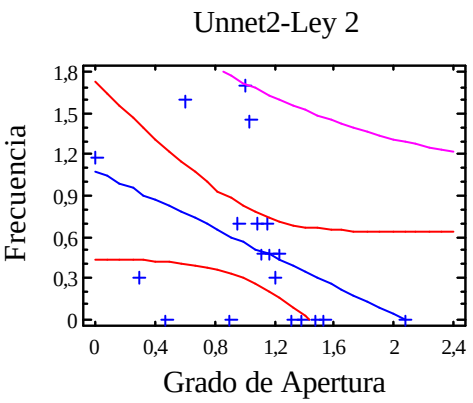


Figura 82: Peor CCA en Ley 2-R2

4.3.3. Tercera recogida.

Ninguno de los dominios presenta un coeficiente de correlación absoluto (CCA) >= 0,95 indicando que esta ley no queda bien representada con este exponente en los dominios españoles. Si tenemos en cuenta que el porcentaje de implantación de la Ley 1 no es demasiado fuerte, al representar esta ley mediante

la frecuencia del grado de apertura, nos indica que los dominios españoles no siguen ninguna pauta en cuanto a esa frecuencia del grado de apertura.

El 26% de los dominios tiene un CCA > = 0,9 y < = 0,95 indicándonos que algunas de las características determinantes en la presencia de esta ley tienen un menor peso en su implantación.

El 74% de los dominios no presenta esta ley y por ello carecen de las características mencionadas anteriormente.

Respecto al valor de la pendiente obtenida por los dominios, está muy por debajo de los valores obtenidos en (Faloutsos, 1999) pero manteniéndose en valores bastante similares, indicándonos que la distribución del exponente *O* en los dominios españoles es bastante similar.

Dominio	Correlacion-R3	Pendiente-R3
Ugr	-0,941964	-1,41422
Unex	-0,926402	-1,70868
Unican	-0,911104	-1,66514
Uva	-0,902202	-1,61409
Ujaen	-0,898017	-1,5793
Urv	-0,896902	-1,34968
Hrc	-0,896771	-1,37242
Esa	-0,895852	-1,51463
Vhebron	-0,888024	-1,70769
Impi	-0,885231	-0,989575
Uc3m	-0,884609	-1,4138
Upco	-0,880793	-1,85661
Fundesco	-0,877541	-1,49609
Us	-0,874052	-1,27625
Usal	-0,823341	-1,26464
Cicyt	-0,817003	-1,20844
Uji	-0,814528	-1,48006
Deusto	-0,807875	-1,40718
Uv	-0,792294	-0,828799
Upsa	-0,779225	-1,14951
Ciemat	-0,715264	-1,06908
Url	-0,706246	-1,2251
Uam	-0,702407	-1,12131
Unnet	-0,68894	-0,863662

Cervantes	-0,585063	-1,0196
Um	-0,48797	-0,766098
Upna	-0,0831257	-0,120009

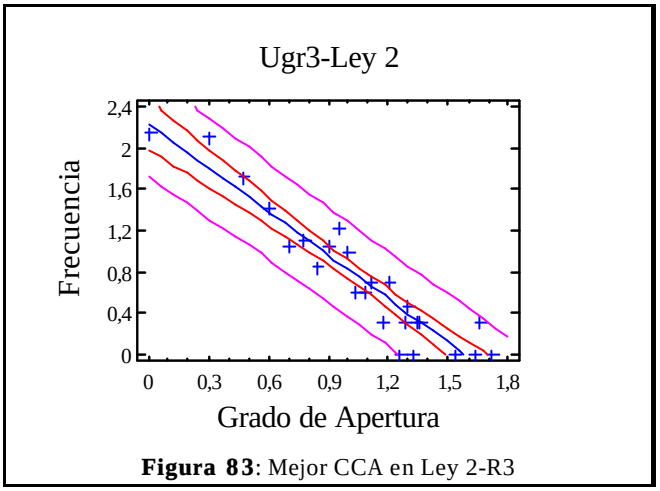


Figura 83: Mejor CCA en Ley 2-R3

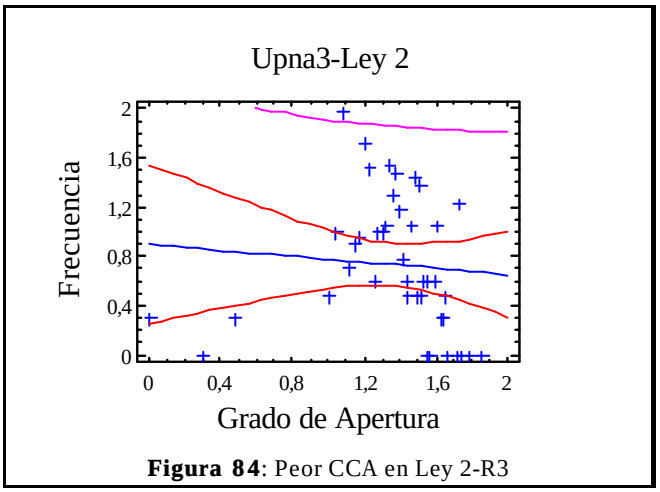


Figura 84: Peor CCA en Ley 2-R3

4.4. Exponente de Hop-plot H (Ley de exponente 3).

Con esta ley se intenta cuantificar la conectividad y la distancia de los nodos de un determinado dominio analizado o de una topología de red concreta (Faloutsos, 1999). Para ello se trabaja con los pares de nodos que tienen un

determinado número de saltos (que están a una distancia determinada). De esta forma para calcular, por ejemplo, los pares de nodos que tienen un salto 5, se tienen en cuenta todos los pares que tienen 5 saltos o menos. De esta forma podremos representar el número de saltos y los pares de nodos que tienen ese número de saltos.

Podemos definir formalmente la Ley 3 de la siguiente forma: el número de pares de nodos, P(h), con h saltos, es proporcional al número de saltos elevado a un exponente H.

$$P(h) a\ h^{H},\ si\ h << d$$

El exponente de salto H, es la pendiente que se obtiene con la representación de los pares de nodos P(h) con h saltos frente al número de saltos en una escala logarítmica.

Este exponente representa la conectividad de los grafos diferenciando eficientemente familias de grafos.

De la aplicación de esta ley, podemos sacar algunos datos muy útiles como la de calcular el diámetro efectivo del Web (Faloutsos, 1999), diferenciándolo de los datos aportados por (Albert, 1999) para calcular dicho diámetro.

El diámetro lo podemos definir como la medida de la distancia más corta existente entre dos nodos del grafo, y ello implica que cuando dos nodos tienen un determinado diámetro existe una alta probabilidad de que exista un enlace entre ellos. Lógicamente, el cálculo del diámetro sobre valores no reales, sino aproximados o fruto de un cálculo ofrecen una visión de la tendencia que tienen ese grafo.

Para (Faloutsos, 1999), dado un grafo de N nodos, E aristas y un exponente de salto H, podemos definir el diámetro efectivo, d_{ef} como:

$$d_{ef} = \left(\frac{N^2}{N + 2E} \right)^{1/H}$$

Para (Albert, 1999) el cálculo del diámetro se realiza de la siguiente forma $d = 0{,}35 + 2{,}06 \log (N)$, siendo N el número de nodos del grafo.

Los trabajos de (Medina, 2000) indican que para que exista la ley de exponente 3 no se precisa de crecimiento exponencial, ni de conectividad preferencial.

Los valores obtenidos en la investigación son los siguientes:

4.4.1. Primera recogida.

El 78% de los dominios presenta un coeficiente de correlación absoluto (CCA) > = 0,95 indicando la presencia de dicha ley.

El 15% de los dominios tiene un CCA > = 0,9 y < = 0,95 indicándonos que referente a la conectividad de los dominios en general se encuentran bastante bien conectados.

Respecto al valor de la pendiente obtenida por los dominios con CCA más alto, es aproximadamente la mitad de los valores obtenidos en (Faloutsos, 1999) pero manteniendo una cierta disparidad en sus valores, indicándones que la distribución del exponente H en los dominios españoles es bastante dispar y la evolución en lo referente a la conectividad no se realiza atendiendo a ningún mecanismo común en su crecimiento y desarrollo.

Dominio	Correlacion-R1	Pendiente-R1
Urv	0,993859	2,19104
Us	0,9905	2,27866
Deusto	0,990448	1,70137
Upsa	0,990129	1,6446
Ugr	0,985818	2,10209
Unex	0,984159	1,89846

Fundesco	0,983999	1,59666
Usal	0,98002	2,61678
Cervantes	0,979944	1,92357
Um	0,978938	2,5976
Unican	0,978463	2,63268
Uam	0,977879	2,09586
Hrc	0,975885	01,57854
Ujaen	0,975118	2,17704
Url	0,974238	1,7281
Uva	0,970537	2,28956
Uji	0,968484	2,98834
Uv	0,961215	2,59437
Impi	0,954001	1,32061
Esa	0,94939	1,78103
Vhebron	0,949035	1,53153
Upna	0,941884	1,42569
Cicyt	0,93957	0,545612
Ciemat	0,93868	2,83895
Upco	0,913236	1,36926
Unnet	0,861108	1,34794
Uc3m	0,831369	0,856736

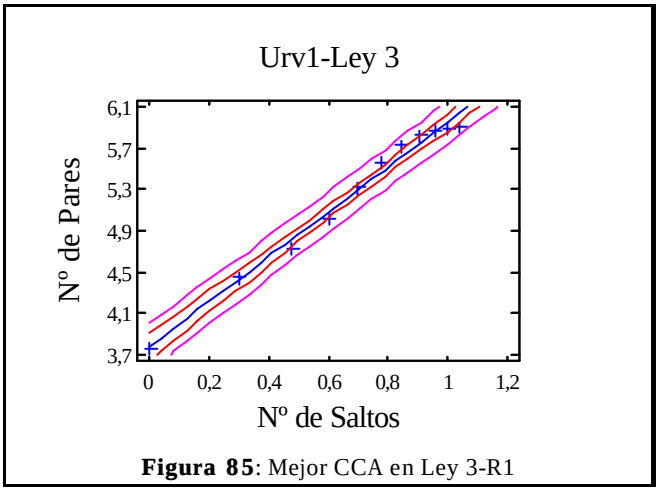


Figura 85: Mejor CCA en Ley 3-R1

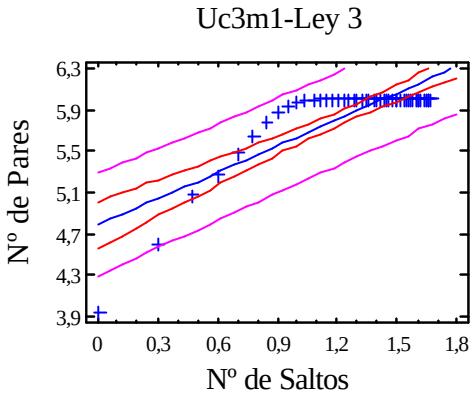


Figura 86: Peor CCA en Ley 3-R1

4.4.2. Segunda recogida.

El 70,4% de los dominios presenta un CCA > = 0,95 indicando la presencia de dicha ley y con un ligero descenso respecto a la primera recogida.

El 26% de los dominios tiene un CCA > = 0,9 y < = 0,95 indicándonos que respecto a la conetividad de los dominios ésta es alta, aumentando ligeramente respecto a la primera recogida.

Dominio	Correlacion-R2	Pendiente-R2
Fundesco	0,990694	1,5554
Deusto	0,986585	1,7458
Urv	0,985434	2,1823
Uc3m	0,984163	2,59607
Upsa	0,98239	1,68783
Unex	0,980754	1,77847
Hrc	0,976774	1,7053
Unican	0,976314	2,39489
Us	0,975319	1,75037
Esa	0,975127	2,19694
Ugr	0,969219	1,72575
Uva	0,965448	2,13478
Uv	0,963086	1,89939

Ujaen	0,962675	2,46153
Impi	0,956917	1,31909
Usal	0,956214	1,80552
Url	0,955108	1,6597
Ciemat	0,954714	3,12833
Uam	0,950127	1,49653
Um	0,942951	2,59668
Vhebron	0,931818	1,17354
Upco	0,928337	1,28221
Cicyt	0,915703	1,51797
Upna	0,914709	1,64536
Cervantes	0,912055	1,6611
Uji	0,903291	1,97733
Unnet	0,888618	1,50578

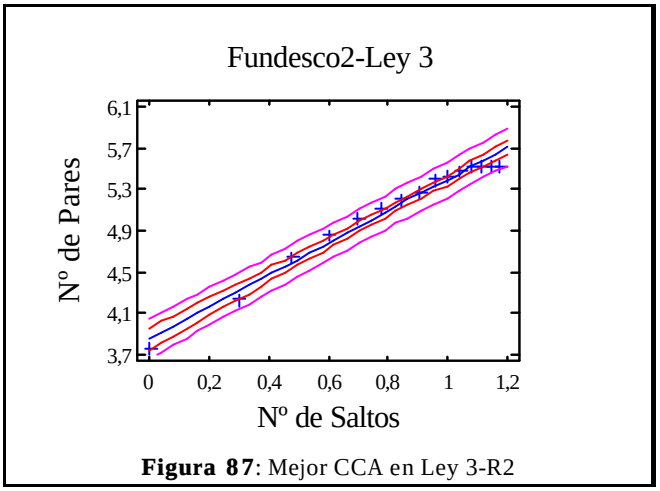


Figura 87: Mejor CCA en Ley 3-R2

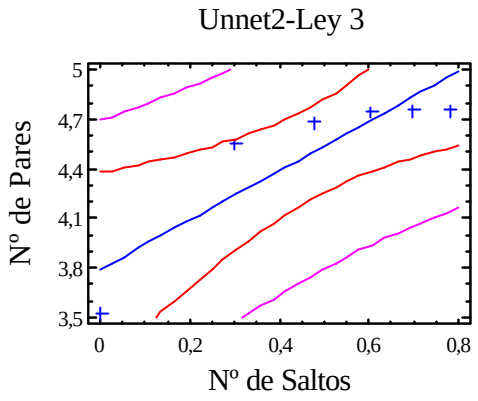


Figura 88: Peor CCA en Ley 3-R2

4.4.3. Tercera recogida.

El 74,1% de los dominios presenta un CCA > = 0,95 indicando la presencia de dicha ley y con un ligero descenso respecto a la primera recogida.

El 18,5% de los dominios tiene un CCA > = 0,9 y < = 0,95 indicándonos que respecto a la conectividad de los dominios ésta es alta, aumentando ligeramente respecto a la primera recogida.

Dominio	Correlacion-R3	Pendiente-R3
Fundesco	0,997406	1,59103
Unex	0,992409	1,72379
Upsa	0,984884	1,60115
Deusto	0,984272	1,97714
Urv	0,983305	2,20191
Vhebron	0,979962	1,78383
Ujaen	0,977379	2,68221
Us	0,973872	1,92584
Hrc	0,973682	1,78565
Esa	0,971914	2,23809
Unican	0,967704	2,25566
Impi	0,96758	1,04231
Uc3m	0,964015	2,33249
Uva	0,963982	2,22337

Url	0,959905	1,47116
Ugr	0,959372	1,63436
Ciemat	0,953393	3,12271
Uam	0,95154	2,00792
Um	0,950308	2,10433
Usal	0,950108	1,82168
Upco	0,934265	1,32022
Upna	0,932643	1,36432
Cicyt	0,921503	1,8362
Uji	0,916719	2,16098
Cervantes	0,912411	1,66401
Unnet	0,873061	1,16077
Uv	0,841902	0,703225

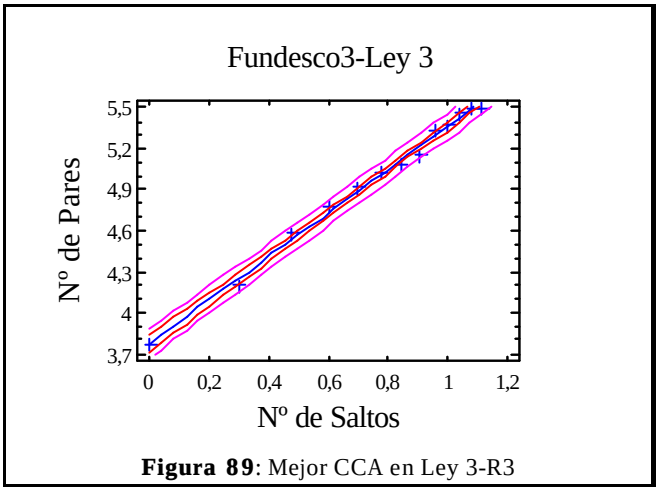


Figura 89: Mejor CCA en Ley 3-R3

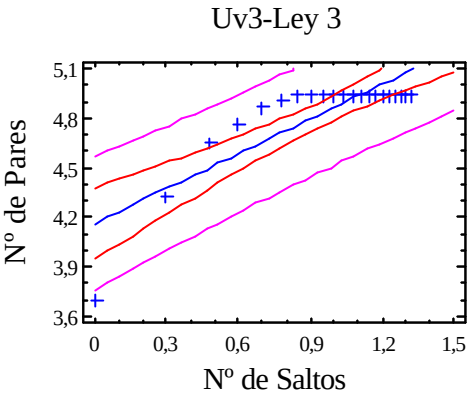


Figura 90: Peor CCA en Ley 3-R3

4.4.4. El diámetro efectivo.

Para analizar los datos del diámetro tenemos los datos reales de la distancia máxima existente en los grafos formados por cada una de las recogidas (y que ya indicamos en la página 116) y los cálculos realizados para cada recogida según el diámetro efectivo de (Faloutsos, 1999) y el diámetro Web de (Albert, 1999), que nos permitirán comparar estos cálculos con los datos reales obtenidos.

Otros trabajos en los que se ha calculado el diámetro Web son los de (Govindan, 1997) que obtiene un diámetro entre 9 y 10 del estudio realizado sobre una recogida en 900 dominios durante 1995. (Albert, 1999) mediante la aplicación de su cálculo del diámetro obtiene un valor de 19 como díametro del Web considerando que el número de nodos del Web es $8x10^8$.

Del estudio de los datos obtenidos podemos indicar a priori que el diámetro Web de Albert no refleja realmente la distancia y las variaciones en la misma que realmente se producen, debido a que su cálculo se realiza solamente a través del número de nodos. En la medida de Faloutsos al tenerse en cuenta el número de nodos, el de aristas y sobre todo el valor del exponente Hop-plot, que mide en cierta medida la conectividad del grafo, refleja mucho mejor las variaciones que se producen en el grafo, respecto a su diámetro.

Por ello en la explicación de los datos vamos a centrar nuestros comentarios en el diámetro efectivo, aunque a modo de comparación se incluyen los datos relativos al diámetro Web.

4.4.4.1. Primera recogida.

Los datos obtenidos nos muestran una enorme variación en la diferencia existente entre la distancia real y el diámetro efectivo de Faloutsos. En general este último valor es superior al dato real, aunque en algunos casos desciende dicho valor. No se aprecia ningún aumento o descenso constante en los datos. Hay que destacar el dominio uc3m con un diámetro real muy alto, que se ve reflejado en el diámetro efectivo. En esta recogida el 63% tiene un diámetro igual o superior a 10.

Dominio	Real	Efectivo	Albert
Cervantes	13	18,09	6,53
Cicyt	3	4,13	3,79
Ciemat	7	8,01	6,53
Deusto	11	19,18	6,24
Esa	10	13,01	5,62
Fundesco	16	25,99	6,38
Hrc	5	5,78	3,76
Impi	15	35,81	5,81
Uam	11	10,82	6,53
Uc3m	47	513,99	6,53
Ugr	14	12,97	6,53
Ujaen	12	13,66	6,53
Uji	6	6,95	6,53
Um	8	7,38	6,53
Unex	15	14,2	6,53
Unican	7	6,81	6,53
Unnet	7	18,29	4,98
Upco	16	48,96	6,34
Upna	17	36,58	6,34
Upsa	10	11,73	5,06
Url	14	20,31	6,53
Urv	11	10,23	6,53
Us	10	8,53	6,53
Usal	9	7,25	6,53
Uv	7	6,6	6,53
Uva	9	10,67	6,53

Vhebron	19	31,83	6,53

4.4.4.2. Segunda recogida.

En esta recogida el 70% de los dominios dispone de un diámetro igual o superior a 10. Un aumento importante se produce en el dominio cicyt, con aumentos importantes en upco y vhebron.

Dominio	Real	Efectivo	Renka
Cervantes	12	34,44	6,53
Cicyt	16	46,81	6,53
Ciemat	6	6,65	6,53
Deusto	15	21,99	6,53
Esa	10	12,04	6,53
Fundesco	15	24,3	6,14
Hrc	7	6,38	4,39
Impi	15	34,23	5,81
Uam	13	28,45	6,53
Uc3m	9	8,28	6,53
Ugr	13	19,62	6,53
Ujaen	9	10,56	6,53
Uji	10	19,12	6,53
Um	8	8,7	6,53
Unex	16	18,41	6,53
Unican	8	9,5	6,53
Unnet	6	13,56	4,95
Upco	30	78,38	6,53
Upna	10	31,92	6,53
Upsa	11	13,63	5,63
Url	15	21,12	6,53
Urv	13	11,55	6,53
Us	14	17,63	6,53
Usal	12	20,59	6,53
Uv	8	12,52	6,53
Uva	11	10,92	6,53
Vhebron	27	76,9	6,34

4.4.4.3. Tercera recogida.

El 74% de los dominios dispone de un diámetro igual o superior a 10, aunque en conjunto se mantienen por debajo de los valores extremos observados en la segunda recogida.

Dominio	Real	Efectivo	Renka
Cervantes	11	34,21	6,53
Cicyt	11	22,01	6,53
Ciemat	6	6,64	6,53
Deusto	11	12,5	6,53
Esa	10	11,26	6,53
Fundesco	13	24,95	6,26
Hrc	8	7,31	4,78
Impi	9	51,54	5,88
Uam	11	18,21	6,53
Uc3m	11	10,7	6,53
Ugr	16	23,34	6,53
Ujaen	8	8,38	6,53
Uji	9	14,69	6,53
Um	7	15,47	6,53
Unex	15	19,77	6,53
Unican	10	10,79	6,53
Unnet	10	31,97	5,16
Upco	25	68,21	6,53
Upna	9	45,88	5,91
Upsa	12	16,8	5,82
Url	20	34,78	6,53
Urv	12	10,1	6,53
Us	14	11,87	6,53
Usal	12	19,09	6,53
Uv	21	566,12	6,53
Uva	10	10,07	6,53
Vhebron	16	18,97	6,53

Un reflejo gráfico de lo comentado anteriormente, lo podemos ver en el siguiente gráfico.

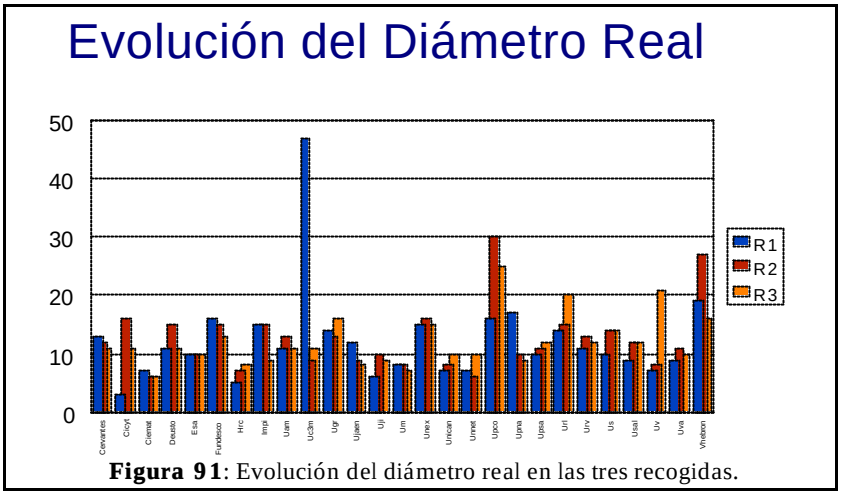


Figura 91: Evolución del diámetro real en las tres recogidas.

Analizando el valor promedio en cada una de las recogidas, podemos ver como quedan solapados los valores promedio de las tres recogidas manteniendo por lo tanto, a pesar del ligero aumento del porcentaje de dominios con mayor diámetro, un valor constante de forma global.

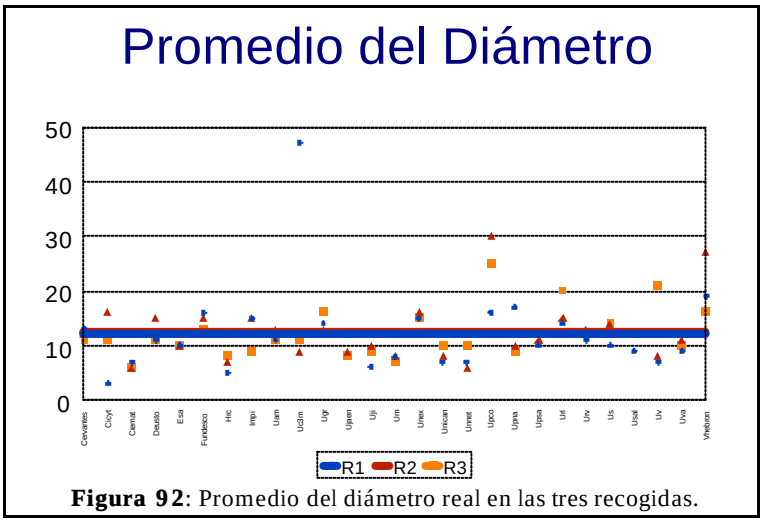


Figura 92: Promedio del diámetro real en las tres recogidas.

4.5. Exponente de valores propios e (Ley de exponente 4).

Los valores propios, l_i de un grafo son proporcionales al orden i, elevado a un exponente, e .

$$l_i \; a \; i^e$$

El exponente de valores propios e es la pendiente que se obtiene al representar los valores propios frente a su orden en una escala logarítmica.

Este exponente también nos permite caracterizar diferentes topologías Web, que además es independiente del crecimiento de dicho Web.

Los valores propios de una matriz están relacionados con algunas propiedades como pueden ser el diámetro, el número de aristas, el número de componentes conectados, el número de rutas existentes que poseen una determinada longitud, que son aspectos fundamentales dentro del análisis de cualquier topología.

4.5.1. Primera recogida.

El 74,1% de los dominios presenta un CCA > = 0,95 indicando la presencia de dicha ley.

El 11,1% de los dominios tiene un CCA > = 0,9 y < = 0,95

Dominio	Correlacion-R1	Pendiente-R1
Uam	-0,995637	-0,580778
Unex	-0,991054	-0,524612
Upna	-0,990444	-0,465705
Ciemat	-0,989461	-0,334667
Urv	-0,98928	-0,508951
Upco	-0,989165	-0,425234
Uva	-0,98862	-0,472291
Us	-0,98773	-0,611047
Cervantes	-0,985731	-0,334146
Ugr	-0,984109	-0,364826

Unican	-0,981465	-0,569937
Usal	-0,980907	-0,346915
Uv	-0,972605	-0,525005
Uc3m	-0,967541	-0,409949
Uji	-0,966729	-0,569792
Unnet	-0,966363	-0,913081
Ujaen	-0,960993	-0,264655
Um	-0,956841	-0,380673
Url	-0,950351	-0,59172
Impi	-0,948321	-0,356504
Vhebron	-0,940907	-0,291075
Fundesco	-0,92905	-0,623535
Esa	-0,925375	-0,748032
Upsa	-0,77804	-6,87538
Hrc	-0,760285	-1,57409
Deusto	-0,540193	-4,01972
Cicyt	0	0

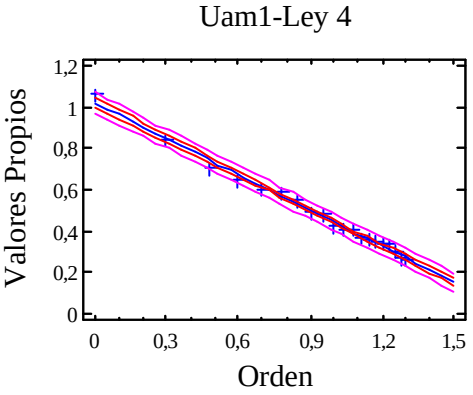


Figura 93: Mejor CCA en Ley 4-R1

4.5.2. Segunda recogida.

El 74,1% de los dominios presenta un CCA > = 0,95 indicando la presencia de dicha ley.

El 18,5% de los dominios tiene un CCA > = 0,9 y < = 0,95

Dominio	Correlacion-R2	Pendiente-R2
Ciemat	-0,995717	-0,353943
Upco	-0,988144	-0,528383
Um	-0,982634	-0,360959
Uji	-0,981273	-0,429645
Fundesco	-0,979797	-0,753611
Us	-0,979652	-0,773672
Uam	-0,979641	-0,690552
Esa	-0,978688	-0,387128
Cervantes	-0,978657	-0,47962
Upna	-0,978044	-0,610918
Usal	-0,977624	-0,352895
Uv	-0,977319	-0,630363
Unican	-0,975925	-0,540369
Uc3m	-0,974374	-0,410312
Unex	-0,972302	-0,413937
Ujaen	-0,972231	-0,302504
Ugr	-0,968238	-0,431812
Url	-0,963041	-0,475689
Unnet	-0,962673	-0,909758
Uva	-0,952758	-0,39241
Urv	-0,933568	-0,415827
Impi	-0,933426	-0,394162
Upsa	-0,928769	-0,816269
Deusto	-0,925788	-0,246105
Vhebron	-0,90901	-0,251878
Cicyt	-0,886904	-0,3935
Hrc	-0,694238	-13,4657

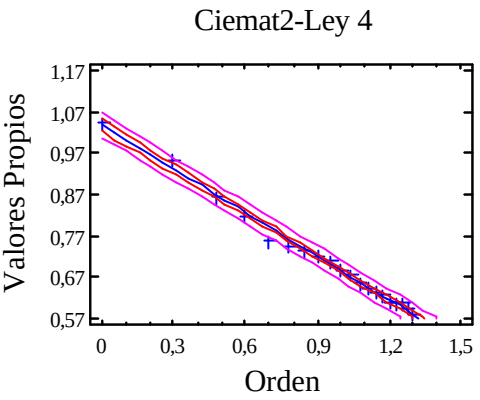


Figura 94: Mejor CCA en Ley 4-R2

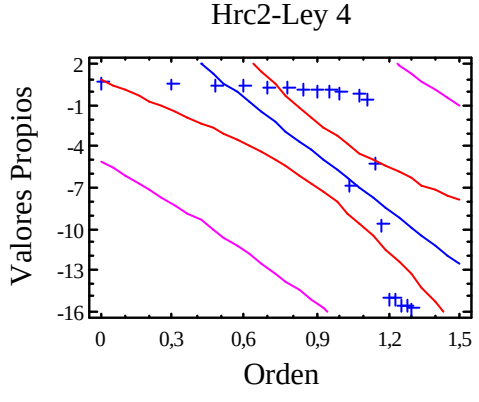


Figura 95: Peor CCA en Ley 4-R2

4.5.3. Tercera recogida.

El 92,6% de los dominios presenta un CCA > = 0,95 indicando la presencia de dicha ley.

El 3,7% de los dominios tiene un CCA > = 0,9 y < = 0,95

Dominio	Correlacion-R3	Pendiente-R3
Ciemat	-0,995096	-0,36087
Uam	-0,991604	-0,431875
Vhebron	-0,989198	-0,191075
Upna	-0,98783	-0,612154
Unican	-0,986785	-0,416983
Uji	-0,985423	-0,410194
Ujaen	-0,980979	-0,305689
Hrc	-0,980699	-0,688144
Urv	-0,979405	-0,456681
Esa	-0,978922	-0,399288
Ugr	-0,977902	-0,448162
Us	-0,976457	-0,804701
Fundesco	-0,976337	-0,785968
Upsa	-0,972795	-0,910666
Upco	-0,97186	-0,538908
Unnet	-0,970794	-0,891163
Uv	-0,969666	-0,607252
Cervantes	-0,968295	-0,38193
Usal	-0,968004	-0,347265
Uva	-0,966635	-0,391754
Deusto	-0,966078	-0,759574
Unex	-0,965407	-0,375573
Url	-0,953631	-0,508187
Uc3m	-0,950885	-0,428071
Um	-0,950305	-0,550019
Cicyt	-0,939667	-0,517305
Impi	-0,791458	-23,5411

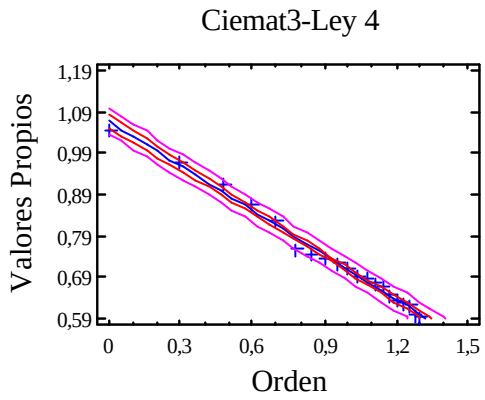


Figura 96: Mejor CCA en Ley 4-R3

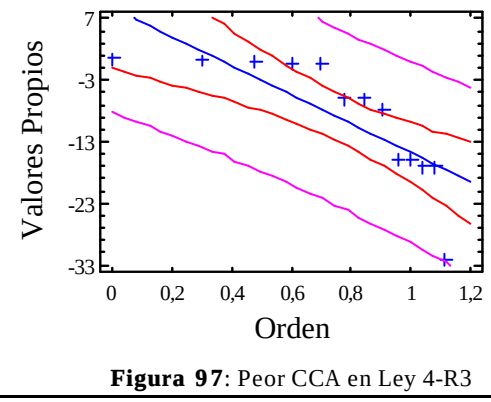


Figura 97: Peor CCA en Ley 4-R3

4.6. Conclusiones.

El estudio de las leyes de exponenciación nos ha permitido un primer acercamiento a las mismas, posibilitando calcular dichos datos para la realidad española. Esto nos permite comparar los datos obtenidos, con los datos de las

investigaciones de Faloutsos y Medina, aportando una nueva realidad al estudio de las mismas.

Respecto a la Ley de exponente 1, se aprecia un empeoramiento en el ajuste de los dominios Web españoles a esta ley, sufriendo peores valores en cada una de las recogidas. A pesar de ello, los dominios que obtienen los valores de CCA más altos y que por lo tanto presentan esta ley, son los que siguen manteniendo estos buenos niveles e incluso mejorándolos. Esto significa que los dominios que tienen un buen diseño topológico siguen manteniendo este buen diseño e incluso lo mejoran. Algunos dominios que se encontraban dentro de CCA por debajo de 0,95, pero por encima de 0,9 en recogidas posteriores son los que han mejorado dicha topología y pasan a estar englobados dentro del grupo de los dominios de CCA superior a 0,95.

Si tenemos en cuenta las teorías de (Medina, 2000) el empeoramiento en los resultados en las sucesivas recogidas tienen también implicaciones respecto al crecimiento exponencial y a la conectividad preferencial, perdiendo en ambos casos importancia, a medida que transcurre el tiempo.

Respecto a la ley de exponente 2, ningún dominio presenta esta ley de forma adecuada en ninguna de las recogidas, pues ninguno tienen un CCA superior a 0,95 en ninguna de las recogidas. Respecto al grupo de dominios que estarían entre 0,9 y 0,95 se aprecia una mejora en la segunda recogida, respecto a la primera, para pasar a la tercera recogida, con peores resultados que en la primera.

Desde el punto de vista del grado de apertura de los diferentes dominios los resultados nos indican que no existen mecanismos de conexión entre los dominios que sigan un patrón establecido.

Respecto a la ley de exponente 3, los niveles de ajuste a la ley son muy buenos en todas las recojidas y se mantienen los mismos dominios, en general dentro de los que poseen mejores resultados.

Desde el punto de vista de la frecuencia de los grados de apertura los resultados son bastante parejos en todas las recogidas. Además al no influir el crecimiento exponencial y la conectividad preferencial en la presencia de esta ley, posiblemente por ello los resultados son mejores que en el caso de las dos leyes

anteriores, que parece que se ven muy influidas por la no presencia de estas dos características en la realidad española.

Dentro de esta ley, analizando la característica del diámetro podemos concluir que el cálculo del diámetro según la ley de exponente 3 suele estar por encima de los datos del diámetro real de los dominios analizados, excepto en unos pocos casos en los que casi coincide o es inferior. Lógicamente, los dominios que mejores resultados han obtenido en la ley de exponente 3, son los dominios que tienen un ajuste mejor al diámetro real existente.

Este diámetro nos da una idea de qué diámetro se puede llegar a alcanzar, según la distribución existente del grado de apertura, indicando sobre todo una tendencia de los dominios.

Respecto a la ley de exponente 4, la mejora se produce en todas las recogidas ajustándose cada vez más a la existencia de esta ley y mostrando que en conjunto los dominios españoles, teniendo en cuenta la distribución de los valores propios, siguen unos patrones más o menos comunes y que englobando las diferentes características a las que se refieren los valores propios los dominios se mantienen de forma bastante similar.

5. Conclusiones.

5.1. Conclusiones.

Para las conclusiones finales del trabajo de investigación, vamos a volver a analizar los objetivos iniciales planteados y haremos las indicaciones oportunas sobre ellos.

Objetivo. Plantear un procedimiento de investigación cibermétrica, empleando para ello un sistema de recogida de datos elaborado por nosotros, que nos permita realizar cualquier uso de cálculo que precisemos y eliminando por ello los inconvenientes que plantean los sistemas que se están utilizando hasta este momento.

Conclusión. La decisión que adoptamos al comenzar este trabajo de investigación en el año 1997, de crear nuestro propio robot para la recogida de datos, se ha mostrado como una herramienta muy eficaz, extremadamente potente y configurable que nos permite realizar cualquier recogida de datos del Web que precisemos, recogiendo la información que deseemos en cada ocasión y que almacena los datos en una base de datos relacional que nos permite el empleo del SQL para la realización de todas las operaciones necesarias en el tratamiento de la información.

Además nos ha permitido sortear las limitaciones de otros sistemas empleados hasta la fecha para la recogida de datos del Web y que ya comentamos oportunamente, e incluso algunos de los indicadores, que presentaban determinados sesgos debido al sistema de recogida de datos han quedado solventados mediante este sistema.

Podemos realizar cualquier procesamiento, sin ningún límite y ofrecer los resultados en ficheros formateados según nuestras necesidades, que pueden ser posteriormente procesados mediante herramientas mucho más sofisticadas para el tratamiento como puede ser Matlab.

Objetivo. Plantear tres posibles líneas de trabajo, que en conjunto ofrezcan una

nueva visión de los estudios cibermétricos y que hasta este momento no se han planteado.

Conclusión. Las tres líneas de investigación analizadas han sido un nuevo planteamiento no abordado hasta este momento, que nos ha permitido valorar facetas que normalmente no se analizaban de forma conjunta. La mayor parte de los trabajos de análisis del Web se han basado en el análisis cuantitativo, posiblemente por los métodos empleados en la recogida de datos, que hacían inviables algunos de los estudios al no poder realizar un cálculo adecuado. Con la metodología de recogida de datos que hemos planteado cualquier estudio es posible y podemos abordar varios criterios en paralelo, que nos ofrecen una visión mucho más amplia del problema.

Objetivo. Plantear un análisis cuantitativo en nuestro estudio que nos permita obtener una serie de indicadores que nos den una medida de la evolución de los dominios Web españoles. Realizaremos un estudio de los análisis planteados hasta el momento y las soluciones que nosotros hemos ofrecido para algunos de los indicadores planteados.

Conclusión. Por un lado está el análisis que hemos denominado cuantitativo, en el que analizamos muchos aspectos diferentes, que nos dan una idea de la evolución del Web, desde el punto de vista de las etiquetas empleadas o de los diferentes tipos de ficheros o del empleo del estándar de exclusión; que también nos permite analizar la estructura hipertexto del Web, valorando la calidad del diseño hipertexto, el factor de impacto e incluso plantear un análisis de cocitas al estilo de la bibliometría clásica.

Los planteamientos abordados, creemos que han sobrepasado en mucho algunas de las tendencias existentes en algunos investigadores del Web, que solamente plantean la realización de un análisis bibliométrico clásico, bien es cierto que sin ofrecer los mecanismos necesarios para hacerlo, teniendo en cuenta las particularidades de la información con la que trabajamos.

Se ha abordado la mayor parte de los aspectos que se están tratando en un análisis cuantitativo y los resultados obtenidos han sido muy adecuados e incluso han superado algunas de las expectativas que manejábamos en un principio. Los resultados han sido muy satisfactorios, aunque para que tengan validez creemos que lo más importante es emplear estos resultados para aplicarlos a recuperación de la información e incluso a minería Web, que se están revelando en este momento como dos opciones muy interesantes de trabajo.

Objetivo.

Estudiaremos el Web como si fuese un grafo, según indican algunas de las teorías y analizaremos desde esa perspectiva algunos cálculos que creemos nos van a permitir analizar la topología de los dominios y ofreciendo unos valores adecuados que nos permitirán analizar la evolución de dichos dominios. Se realizará la comparación de algunos cálculos que en teoría miden lo mismo e intentaremos observar si esto es realmente así o por el contrario tienen algún comportamiento diferenciador.

Conclusión.

El considerar el Web como un grafo no es una teoría novedosa, pero muchos de los trabajos de análisis del Web obvian ete planteamiento, posiblemente por las dificultades que plantea en la recogida de datos y su posterior tratamiento, precisando de equipos y tiempos de cálculo en general elevados.

Con este planteamiento hemos podido realizar el cálculo de algunos indicadores que nos han permitido valorar la estructura del grafo desde un punto de vista topológico y que ofrece una nueva visión en el análisis del mismo. Nos hemos centrado en unos pocos índices de los muchos que se pueden utilizar. Abrimos una puerta de trabajo y ofrecemos datos para el estudio del Web español que no se habían realizado todavía y que ofrecen unos datos reales de la situación del Web en tres momentos distintos y nos da una idea de como ha evolucionado en el tiempo. Estos planteamientos tienen aplicación directa en recuperación de la información, pudiendo resultar de enorme interés.

Objetivo. Abordaremos una nueva tendencia de estudio muy reciente, denominada leyes de exponenciación, e intentaremos aplicar dichas leyes a los dominios objeto del estudio analizando las características que estos dominios puedan plantear en función de los cálculos realizados.

Conclusión. Después de finalizar el trabajo en esta línea de investigación, aunque los resultados que ofrecemos son totalmente novedosos y sirven como punto de referencia con algunos de los pocos trabajos existentes de este tipo, los resultados obtenidos posiblemente sean los que menos satisfechos nos han dejado. Podemos valorar diferentes aspectos topológicos del Web, con un planteamiento totalmente novedoso, pero que con toda seguridad requiere de nuevos trabajos que asienten tanto el aspecto teórico como la valoración que se puede realizar de los resultados. Los trabajos actuales en su mayor parte se han realizado en ensayos de laboratorio y en simulaciones.

A pesar de todo, intuimos, si la intuición puede tener cabida en un trabajo de investigación, que puede ser una vía de estudio interesante.

Objetivo. Debemos valorar si la metodología utilizada y las líneas de investigación que se abordan han sido adecuadas y obtenemos resultados válidos o resultados que puedan ser prometedores, pero que precisan de una mayor investigación y desarrollo.

Conclusión. Finalmente, creemos que los planteamientos son correctos y han dado resultados de interés en este antiguo y nuevo campo de trabajo. Las leyes de exponenciación requieren de un mayor asentamiento y de un mayor trabajo que conformen esta nueva teoría.

5.2. Líneas de trabajo futuro.

Finalmente vamos a indicar algunas de las posibles líneas de trabajo, algunas de las cuales ya se encuentran en pleno proceso de asentamiento.

De enorme interés son los estudios relacionados con la recuperación de información en el Web, que con diferentes planteamientos y teorías se está abordando en algunos casos. Creemos que en el mundo de la Documentación uno de los aspectos más importantes es el de recuperación de la información y el Web es un elemento más dentro del sistema documental y que como hemos visto con anterioridad ha adquirido enorme importancia.

Además el gran volumen de información que posee el Web, hace necesario plantear mecanismos que faciliten esta tarea y ofrezcan resultados adecuados.

Algunos trabajos, en los que se dan planteamientos relacionados con la recuperación de información, aunque por supuesto hay más serían los de (Goffinet, 1998), (Golovchinsky, 1997), (Hartman, 1997), (He, 1998), (Koehler, 1999c), (Mukherjea, 1997), (Willet, 1981), (Zhang, 1999), (Zhang, 1999b).

Algunos trabajos en este sentido han sido abordados ya por nuestro grupo de trabajo como se puede ver en (Figuerola, 1998), (Alonso, 1999).

6. Bibliografía.

(Abraham, 1997) ABRAHAM, R. Webometry: measuring the complexity of the World Wide Web. [en línea]. 1997 [Citado: Noviembre 2000]. Disponible en Internet: http://thales.vismath.org/webometry/articles/vienna.html

(Adamic, 1999) ADAMIC, L. A. The Small World Web. *Proceedings of ECDL'99*, p. 443-452.

(Adell, 1994) ADELL, J. y BELLVER, C. Hipermedia distribuido en el Mac: el proyecto World-Wide Web. *I Congreso Universidad y Macintosh*, (UNED, Madrid, Septiembre de 1994).

(Adell, 1994b) ADELL, J. y BELLVER, C. La Internet como telaraña: el World-Wide Web. [en línea]. 1994 [Citado: Septiembre 1999]. Disponible en Internet: http://www.uv.es/~biblios/mei3/Web022.html

(Aguillo, 1996) AGUILLO, I. F. A preliminary Approach to Citation Phenomena in the World Wide Web. *Presented to Signatures of knowledge societies. EASST/4S Conference*, (University of Bielefeld, 10-13 Octiber 1996).

(Aguillo, 1998) AGUILLO, I. F. Hacia un concepto documental de sede Web. *El profesional de la información*, 1998, Vol. 7, No. 1-2, p. 45-46.

(Aguillo, 2000) AGUILLO, I. F. Contenidos de I+D en Internet: Mitos y leyendas. *Mundo Científico*, 2000, No. 211, p. 22-25.

(Aguillo, 2000b) AGUILLO, I. F. Indicadores hacia una evaluación no objetiva (cuantitativa) de sedes Web . *Jornadas Españolas de Documentación*, 2000, Vol. 7, p. 233-248.

(Albert, 1999) ALBERT, R., JEONG, H. y BARABÁSI, A.-L. The Diameter of the World-Wide Web. *Nature*, 1999, Vol. 401, p. 130-131.
Url: http://xxx.lang.gov/abs/cond-mat/9907038

(Almind, 1997) ALMIND, T. C. y INGWERSEN, P. Informetric analyses on the world wide Web: methodological approaches to 'webometrics'. *Journal of Documentation*, September 1997, Vol. 53, No. 4, p. 404-426.

(Alonso, 1997) ALONSO BERROCAL, J. L. *Herramienta software para el análisis de la documentación Web: rastreo de dominios, estudio de etiquetas, tipología de ficheros, evolución de los enlaces.* Salamanca: Universidad de Salamanca, Facultad de Traducción y Documentación, 1997.

(Alonso, 1999) ALONSO BERROCAL, J. L., FIGUEROLA, C. G. y ZAZO RODRÍGUEZ. ÁNGEL FRANCISCO. Representación de páginas Web a través de sus enlaces y su aplicación a la Recuperación de Información. *Scire. Representación y Organización del Conocimiento*, 1999, Vol. 5, No. 2, p. 91-98.

(Arellano, 1999) ARELLANO PARDO, C., RODRÍGUEZ MATEOS, D., NOGALES FLORES, J. T. y HERNÁNDEZ PÉREZ, T. Análisis de estructura de sitios Web: el caso de las bibliotecas universitarias andaluzas. *2as. Jornadas Andaluzas de Documentación, JADOC'99*, (Granada, 1999), p. 39-50.

(Arnzen, 1996) ARNZEN, M. A. Cyber citations: Documenting Internet sources presents some thorny problems. *Internet World*, 1996, Vol. 7, No. 9, p. 72-74.

(ARPANET, 1991)*ARPANET, the Defense Data Network, and Internet*. Encyclopedia of Communications, Volume 1. Editors Fritz Froehlich y Allen Kent. New York: Marcel Dekker, 1991.

(Balasubramanian, 1994) BALASUBRAMANIAN, V. Hypermedia Issues and Applications: A State-of-the-Art Review. *Graduate School of Management, Rutgers University, Newark, New Jersey, 1994*.

(Bar, 2000) BAR-ILAN, J. The Web as an information Source on Informetrics? A content analysis. *Journal of the American Society for Information Science*, 2000, Vol. 51, No. 5, p. 432-443.

(Baran, 1964) BARAN, P. On Distributed Communications Networks. *IEEE Trans. Comm. Systems*, March 1964.

(Bauwens, 1996) BAUWENS, M. Knowledge transfer in cyberspace: A model for future business practices. *FID News Bulletin*, 1996, Vol. 46, No. 1/2, p. 46-54.

(Bayer, 1990) BAYER, A. E., SMART, J. C. y MCLAUHLIN, G. W. Mapping Intellectual Structure of a Scientific Subfield through Author Cocitations. *Journal of the American Society for Information Science*, September 1990, Vol. 41, No. 6, p. 444-452.

(Bergman, 2000) BERGMAN, M. K. The deep Web: Surfacing Hidden Value. [en línea]. 1996 [Citado: Septiembre 2000]. Disponible en Internet: White Paper, BrightPlanet.com LLC, http://www.completeplanet.com/Tutorials/DeepWeb/index.asp

(Berners-Lee, 1993) BERNERS-LEE, T. y CONOLLY, D. Hypertext Markup Language (HTML): A Representation of Textual Information and MetaInformation for Retrieval

and Interchange. [en línea]. Internet Draft, IIIR Working Group, June 1993 [Citado: Marzo 1999]. Disponible en Internet: http://info.cern.ch/hypertext/WWW/MarkUp/HTML.html

(Berners-Lee, 1993b) BERNERS-LEE, T. Hypertext Transfer Protocol: A Stateless Search, retrieve and manipulation protocol. [en línea]. Internet Draft, 1993 [Citado: Marzo 1999]. Disponible en Internet: ftp://nic.switch.ch/mirror/internet-draft/draft-ietf-iiir-http-00.ps

(Berners-Lee, 1994) BERNERS-LEE, T. y. o. The World Wide Web. *Communications of the ACM*, 1994, Vol. 37, No. 8, p. 76-82.

(Berners-Lee, 1994b) BERNERS-LEE, T., MASINTER, L. y MCCAHILL, M. Uniform Resource Locators (URL), RFC 1738. [en línea]. December 1994 [Citado: Septiembre 1999].

(Bernstein, 1992) BERNSTEIN, M. Contours of Constructive Hypertexts. *Proccedings of ACM ECHT CONFERENCE*, (Milano, 30 Noviembre-4 Diciembre de 1992), p. 161-170.

(Bharat, 1998) BHARAT, K. y BRODER, A. A tecnique for measuring the relative size and overlap of public Web serach engines. *Proc. of the Seventh WWW Conference*, (Brisbane, Australia, 1998).

(Bharat, 1998b) BHARAT, K. y HENZINGER, M. R. Improved algorithms for topic distillation in a hyperlinked environment. *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information retrieval*, (1998), p. 104-111.

(Bossy, 1995) BOSSY, M. J. The last of the Litter: "Netometrics". *Solaris*, 1995, No. 2. Url: http://www.info.unicaen.fr/bnum/jelec/Solaris/d02/2bossy.html

(Botafogo, 1991) BOTAFOGO, R. A. y SHNEIDERMAN, B. Identifying aggregates in Hypertext structures. *Proceedings of Hypertext'91*, (Diciembre de 1991), p. 63-74.

(Botafogo, 1992) BOTAFOGO, R. A., RIVLIN, E. y SHNEIDERMAN, B. Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics. *ACM Transactions on Information Systems*, April 1992, Vol. 10, No. 2, p. 142-180.

(Boutell, 1994) BOUTELL, Th. Frequently asked questions about World-Wide Web. [en línea]. 1994 [Citado: Septiembre 1999]. Disponible en Internet: ftp://rtf.mit.edu/usenet/news.answers/www-faq

(Braun, 1985) BRAUN, T. y. o. *Scientometric indicators. A 32 country comparative evaluation of publishing performance and citation impact*. Singapore: World Scientific, 1985.

(Bray, 1996) BRAY, T. Measuring the Web. *Fifth International World Wide Web Conference*, (Paris, France, 6-10 May 1996).
Url: http://www5conf.inria.fr/fich_html/papers/P9/Overview.html

(Brin, 1998) BRIN, S. y PAGE, L. The anatomy of a large-scale hypertextual Web search engine. *Proc. 7th. WWW conference*, (Brisbane, Australia, 14-18 April 1998).
Url: http://www7.scu.edu.au/

(British, 1976) BRITISH STANDARDS INSTITUTION. *Glossary of documentation terms, Published under the authority of the Executive Board on 30 Novembre 1976, prepared under the direction of the Documentation Standards Committeee: 7.*

(Broder, 2000) BRODER, A., KUMAR, R., MAGHOUL, F., RAGHAVAN, P., RAJAGOPALAN, S., STATA, R., TOMKINS, A. y WIENER, J. Graph structure in the Web. *9th. International World Wide Web Conference*, (Amsterdam, May 15 - 19, 2000).
Url: http://www.almaden.ibm.com/cs/k53/www9.final/

(Bush, 1945) BUSH, V. As We May Think. *Atlantic Montly*, 1945, Vol. 176, No. 1, p. 101-108.

(Callon, 1995) CALLON, M., COURTIAL, J.-P. y PENAN, H. *Cienciometría. La medición de la actividad científica: de la bibliometría a la vigilancia tecnológica.* Gijón: ediciones Trea, 1995.

(Campbell, 1896) CAMPBELL, F. *Theory of the National and international Bibliography*. London: Library Bureau, 1896 .

(Carriere, 1997) CARRIERE, J. y KAZMAN, R. Webquery: searching and visualizing the Web through connectivity. *Sixth international World Wide Web conference*, (Santa Clara, California, USA, April 7-11, 1997).

(Castillo, 1999) CASTILLO BLASCO, L., MARTÍNEZ DE PABLOS, M. J. y SERVER, G. Evaluación de la información contenida en seis sedes Web de las escuelas universitarias y facultades de biblioteconomía y documentación españolas. *Revista Española de Documentación Científica* , 1999, Vol. 23 , No. 3, p. 325-332.

(Cawkell, 1976) CAWKELL, A. E. Understanding science by analyzing its literature. *The Information Scientist*, 1976, Vol. 10, No. 1, p. 3-10.

(Cerf, 1974) CERF, V. G. y KAHN, R. E. A protocol for packet network interconnection. *IEEE Trans. Comm. Tech*, May 1974, Vol. COM-22. V.5, p. 627-641.

(Cerf, 1993) CERF, V. G. *How the Internet Came to be. En: Aboba, Bernard. The Online User's Encyclopedia.* Addison-Wesley, 1993.

(Chakrabarti, 1998) CHAKRABARTI, S., DOM, B., RAGHAVAN, P., RAJAGOPALAN, S., GIBSON, D. y KLEINBERG, J. Automatic resource compilation by analyzing hyperlink structure and associated text. *Proc. 7th International World Wide Web Conference*, (1998).
Url: http://decweb.ethz.ch/WWW7/1898/com1898.htm

(Chakrabarti, 1998b) CHAKRABARTI, S. y DOM, B. I. P. Enhanced hypertext categorization using hyperlinks. *Proceedings ACM SIGMOD*, (1998).

(Chakrabarti, 1999) CHAKRABARTI, S., DOM, B., GIBSON, D. y KLEINBERG, J. Mining the link structure of the World Wide Web. *IEEE Computer*, August 1999.

(Chen, 1997) CHEN, C. Structuring and Visualising the WWW by Generalised Similarity Analysis. *Proceedings of Hypertext'97*, (Southampton, UK, 1997), p. 177-186.

(Ciolek, 1997) CIOLEK, T. M. The size, content and geography of asian cyberspace: an initial measurement. [en línea]. 1997 [Citado: Mayo 1998]. Disponible en Internet: http://www.ciolek.com/PAPERS/AsianCyberspace-97.html

(Clarke, 1997) CLARKE, S. J. y WILLETT, P. Estimating the recall performance of Web search engines. *Aslib Proceedings*, July 1997-August 1997, Vol. 49, No. 7, p. 184-189.

(Clausen, 1996) CLAUSEN, H. Looking for the information neddle in the Internet haystack. *Proccedings of the 20th Online Information Meeting, Learned Information*, (Oxford, 1996), p. 115-123.

(Codina, 2000) CODINA, L. Evaluación de recursos digitales en línea: conceptos, indicadores y métodos. *Revista Española de Documentación Científica*, 2000, Vol. 23, No. 1, p. 9-44.

(Coffman, 1998) COFFMAN, K. G. y ODLYZKO, A. The size and growth rate of the internet. *First Monday*, 1998, Vol. 3, No. 10.

(Cole, 1917) COLE, F. J. y EALES, N. B. The history of comparative anatomy: part 1: a statistical analisys os the literature. *Science Progress*, 1917, Vol. 11, p. 578-596.

(Crocker, 1969) CROCKER, S. *RFC001 Host software.*

(Cronin, 1996) CRONIN, B. y MCKIM, G. Science and scholarship on then World Wide Web: a North American Perspective. *Journal of Documentation*, 1996, Vol. 52, No. 2, p. 163-171.

(Cui, 1999) CUI, L. Rating Health Eeb sites using the principles of Citation Analisys: a Bibliometric Approach. *Journal of Medical Internet Research*, 1999, Vol. 1 (1), p. e4. Url: http://www.jmir.org/1999/1/e4/index.htm

(Dahal, 1999) DAHAL, T. M. Cybermetics: The use and implications for Scientometrics and Bibliometrics; A study for Developing Science & Technology Information System in Nepal. *IIIrd National Conference on Science & Technology*, (March 8-11, 1999. Royal Nepal Academy of Science and Technology (RONAST)).
Url: http://www.panasia.org.sg/nepalnet/ronast/cyber.html

(Dean, 1999) DEAN, J. y HENZINGER, M. R. Finding related pages in the World Wide Web. *Computer Networks* , 1999, Vol. 31, No. 11-16, p. 1467-1479.
Url: http://citeseer.nj.nec.com/dean99finding.html

(Diodato, 1994) DIODATO, V. *Dictionary of bibliometrics*. New York: The Haworth Press, 1994.

(Egghe, 1990) EGGHE, L. y ROUSSEAU, R. *Introduction to informetrics. Quantitative methods in library, documentation and information science*. Amsterdam: Elsevier Science publishers, 1990.

(Ellis, 1994) ELLIS, D., FURNER-HINES, J. y WILLETT, P. On the creation of hypertext links in full-text documents: measurement of inter-linker consistency. *Journal of Documentation*, June 1994, Vol. 50, No. 2, p. 67-98.

(Fairthorne, 1969) FAIRTHORNE, R. A. Empirical hyperbolic distributions (Bradford-Zipf-Mandelbrot) for bibliometric description and prediction. *Journal of Documentation*, 1969, Vol. 25, No. 4, p. 319-343.

(Faloutsos, 1999) FALOUTSOS, M., FALOUTSOS, P. y FALOUTSOS, C. On power-law relationships of the internet topology. *ACM SIGCOMM*, (Cambridge, MA, September 1999), p. 251-262.

(Figuerola, 1998) FIGUEROLA, C. G., ALONSO BERROCAL, J. L. y ZAZO RODRÍGUEZ, Á. F. Nuevos puntos de vista en la Recuperación de Información en el Web. *Jornadas Españolas de Documentación*, 1998, Vol. 6, p. 273-280.

(Fonseca, 1973) FONSECA, E. Bibliografia estatistica e bibliometria: Uma reivindicacao de prioridades. *Ciencia da Informacao*, 1973, Vol. 2, No. 1, p. 5-7.

(Garfield, 1976) GARFIELD, E. A bibliometric analysis of references. *Journal Citation Reports*, (Philadelphia, 1976).

(Garfield, 1977) GARFIELD, E. *Historiographs, librarianship and the history of science En: Garfield, E. ed. Essays of an information scientist.* Philadelphia: ISI Press, 1977.

(Garfield, 1978) GARFIELD. E Y OTROS. *Citation data as science indicators. EN: Elkana, Y. ed. toward a metric od science: the advent os science indicators.* New York: John Wiley, 1978.

(Garfield, 1979) GARFIELD, E. Is citation analisys a legitimate evaluation tool? *Scientometrics*, 1979, Vol. 1, No. 4, p. 359-375.

(Garfield, 1979b) GARFIELD, E. *Citation indexing its theory and application in science, technology, and humanities.* New York: John Wiley, 1979.

(Garfield, 1980) GARFIELD, E. *ABCs of cluster mapping: part 1 most active fields in the life sciences in 1978.* EN: Essays of an Information Scientist Ed. by E. Garfield. Philadelphia: ISI Press, 1980.

(Garfield, 1995) GARFIELD, E. New International Professional Society Signals The Maturing of Scientometrics And Informetrics. [en línea]. 1995 [Citado: Febrero 1999]. Disponible en Internet: http://www.the-scientist.library.upenn.edu/yr1995/august/issi_950821.html

(Gibson, 1984) GIBSON, W. *Neuromancer.* New York: Ace Books, 1984.

(Gibson, 1998) GIBSON, D., KLEINBERG, J. y RAGHAVAN, P. Inferring Web communities from link topology. *Proc. 9th ACM Conference on Hypertext and Hypermedia*, (1998).

(Goffinet, 1998) GOFFINET, L. y NOIRHOMME-FRAITURE, M. Automatic Hypertext Link Generation based on Similarity Measures between Documents. [en línea]. 1998 [Citado: Abril 1998]. Disponible en Internet: http://www.info.fundp.ac.be/~lgo/Hypertext/semantic_links.html

(Gollogley, 1997) GOLLOGLEY, G. y SMEATON ALAN F. Assisting the Hypertext Authoring Process with Topology Metrics and Information Retrieval. *Working Papers*, (1997).
Url: ftp://ftp.compapp.dcu.ie/pub/w-papers/1997/CA2897.ps.Z

(Golovchinsky, 1997) GOLOVCHINSKY, G. What the Query Told the Link: The Integration of Hypertext and Information Retrieval. *Proceedings of Hypertext'97*, (Southampton, UK, 1997), p. 67-74.

(Govindan, 1997) GOVINDAN, R. y REDDY, A. An analysis of internet interdomain topology and route stability. *Proc. IEEE INFOCOM*, (Kobe, Japan, April 7-11, 1997).

(Gray, 1995) GRAY, M. Measuring the growth of the Web. [en línea]. 1995 [Citado: Octubre 1999]. Disponible en Internet: http://www.mit.edu/people/mkgray/growth/

(Gray, 1996) GRAY, M. Internet Statistics: Growth and usage of the Web and the Internet. [en línea]. 1996 [Citado: Noviembre 1999]. Disponible en Internet: http://www.mit.edu/people/mkgray/net/

(Harary, 1965) HARARY, F., NORMAN, R. Z. y CARTWRIGHT, D. *Structural models: an introduction to the theory of directed graphs.* New York: Wiley, 1965.

(Harary, 1969) HARARY, F. *Graph Theory*. Reading, MA: Adison Wesley, 1969.

(Hardy, 1993) HARDY, H. The History of the Net. [en línea]. Master Thesis, School of Communications, Grand Valley State University [Citado: Septiembre 1999]. Disponible en Internet: http://www.ocean.ic.net/ftp/doc/nethist.html

(Hardy, 1996) HARDY, I. The Evolution of ARPANET email. [en línea]. History Thesis, UC Berkeley [Citado: Septiembre 1999]. Disponible en Internet: http://www.ifla.org/documents/internet/hari1.txt

(Harter, 1996) HARTER, S. P. The Impact of Electronic Journals on Scholarly Communication: A Citation Analysis. *Public Access Computer Systems Review*, 1996, Vol. 7 (5).
Url: http://info.lib.uh.edu/pr/v7/n5/hart7n5.html

(Harter, 1996b) HARTER, S. P. y HAK, J. K. Electronic Journals and scholarly communication: a citation and reference study. *Proceedings of the ASIS Midyear Meeting*, (San Diego, CA, May, 1996), p. 299-315.
Url: http://ezinfo.ucs.indiana.edu/~harter/harter-asis96midyear.html

(Hartman, 1997) HARTMAN, J. H., PROEBSTING, T. A. y SUNDARAM, R. Index-based hyperlinks. *Sixth international World Wide Web conference* , (Santa Clara, California, USA, April 7-11, 1997).
Url: http://atlanta.cs.nchu.edu.tw/www/PAPER248.html

(Hauben, 1995) HAUBEN, R. y HAUBEN, M. The Netizens and the Wonderful World of the Net. [en línea]. 1995 [Citado: Septiembre 1999]. Disponible en Internet: http://www.columbia.edu/~hauben/netbook/

(Hawkins, 1977) HAWKINS, D. T. Unconventional uses of on-line information retrieval systems: on-line bibliometrics studies. *Journal of the American Society for Information Science*, 1977, Vol. 28, No. 1, p. 13-18.

(Hayes, 2000) HAYES, B. Graph theory in practice: part 1. *A reprint from American Scientist*, January 2000-February 2000, Vol. 88, No. 1, p. 9-13.
Url: http://www.sigmaxi.org/amsci/issues/comsci00/compsci2000-01.ps.gz

(Haynes, 1995) HAYNES, C. *How to succeed in cyberspace*. London: Aslib, 1995.

(He, 1998) HE, S. Concept similarity and conceptual information alteration via English-to-Chinese and Chinese-to-English translation of medical article titles. *Journal of the American Society for Information Science*, 1998, Vol. 49, No. 2, p. 169-175.

(Hobbes, 2000) HOBBES ZAKON, R. Hobbes' Internet Timeline v5.2. [en línea]. 2000 [Citado: Febrero 2000]. Disponible en Internet: http://www.zakon.org/robert/internet/timeline/

(Huberman, 1999) HUBERMAN, B. A. y ADAMIC, L. A. Evolutionary dynamics of the World Wide Web. *Tech. Rep., Xeros Palo Alto Reserach Center*, (February, 1999).

(Huberman, 1999b) HUBERMAN, B. A. y ADAMIC, L. A. Growth dynamics of the World-Wide Web. *Nature*, 1999, Vol. 40, p. 450-457.

(Hulme, 1923) HULME, E. W. *Statistical bibliography in relation to the growth of modern civilization*. London: Grafton, 1923.

(Ingwersen, 1998) INGWERSEN, P. The calculation of Web impact factors. *Journal of Documentation*, March 1998, Vol. 54, No. 2, p. 236-243.

(ISC, 2000) INTERNET SOFTWARE CONSORTIUM (ISC). Internet Host Count. [en línea]. 2000 [Citado: Febrero 2000]. Disponible en Internet: www.isc.org

(ISO 8879:1986) ISO 8879:1986. *SGML*. 1986.

(Kahle, 1989) KAHLE, B. Wide Area Information Servers Concepts. [en línea]. 1989 [Citado: Septiembre 1999]. Disponible en Internet: ftp://ftp.wais.com/pub/wais-inc-doc/wais-concepts.txt

(Kahn, 1972) KAHN, R. Communications Principles for Operating Systems. *Internal BBN memorandum*, January 1972.

(Khan, 1998) KHAN, K. y LOCATIS, C. Searching through cyberspace: the effects of link display and link density on information retrieval from hypertext on the World Wide Web. *Journal of the American Society for Information Science*, 1998, Vol. 49, No. 2, p. 176-182.

(Kleinberg, 1999) KLEINBERG, J. M. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 1999, p. 668-677.

(Kleinberg, 1999b) KLEINBERG, J. M., KUMAR, R. y RAGHAVAN, P. The Web as a graph: measurements, models, and methods. *Proceedings of the Fifth Annual International Computing and Combinatorics Conference*, (1999).

(Kleinrock, 1961) KLEINROCK, L. Information Flow in Large Communication Nets. *RLE Quarterly Progress Report*, July 1961.

(Koehler, 1999) KOEHLER, W. C. An analysis of Web page and Web site constancy and permanence. *Journal of the American Society for Information Science*, 1999, Vol. 50, No. 2, p. 162-180.

(Koehler, 1999b) KOEHLER, W. C. Digital libraries and World Wide Web sites and page persistence. *Information Research*, June 1999, Vol. 4, No. 4.
Url: http://www.shef.ac.uk/~is/publications/infres/paper60.html

(Koehler, 1999c) KOEHLER, W. C. Classifying Web sites and Web pages: the use of metrics and URL characteristics as markers. *Journal of Librarianship and Information Science*, March 1999, Vol. 31, No. 1, p. 21-31.

(Kolb, 1997) KOLB, D. Scholarly Hypertext: self-represented complexity. *Hypertext'97*, (Southampton, UK, 1997), p. 29-37.

(Kollar, 1996) KOLLAR, C., LEAVITT, J. y MAULDIN, M. Robot exclusion standard revisited. [en línea]. 1996 [Citado: Febrero 1999]. Disponible en Internet: http://kollar.com/robots.html

(Koster) KOSTER, M. HTML author's guide to the robots META tag. [en línea].[Citado: Diciembre 1998]. Disponible en Internet:
http://info.webcrawler.com/mak/projects/robots/meta-user.html

(Koster, 1993) KOSTER, M. Guidelines for robot writers. [en línea]. 1993 [Citado: Diciembre 1998]. Disponible en Internet:
http://info.webcrawler.com/mak/projects/robots/guidelines.html

(Koster, 1994) KOSTER, M. A Standard for Robot Exclusion. [en línea]. 1994 [Citado: Diciembre 1998]. Disponible en Internet: http://info.webcrawler.com/mak/projects/robots/norobots.html

(Koster, 1995) KOSTER, M. Robots in the Web: threat or treat? [en línea]. 1995 [Citado: Diciembre 1998]. Disponible en Internet: http://info.webcrawler.com/mak/projects/robots/threat-or-treat.html

(Koster, 1996) KOSTER, M. A Method for Web Robots Control. [en línea]. 1996 [Citado: Diciembre 1998]. Disponible en Internet: http://info.webcrawler.com/mak/projects/robots/norobots-rfc.html

(Koster, 1996a) KOSTER, M. Evaluation of the standard for robots exclusion. [en línea]. 1996 [Citado: Diciembre 1998]. Disponible en Internet: http://info.webcrawler.com/mak/projects/robots/eval.html

(Kulikowski, 1999) KULIKOWSKI, S. I. A Timeline of Network History. [en línea].[Citado: Septiembre 1999]. Disponible en Internet: stankuli@uwf.bitnet

(Kumar, 1999) KUMAR, R., RAGHAVAN, P., RAJAGOPALAN, S. y TOMKINS, A. Extracting large-scale knowledge bases from the Web. *Proceedings of the 25th VLDB Conference*, (Edinburgh, 1999).

(Kumar, 1999b) KUMAR, R., RAGHAVAN, P., RAJAGOPALAN, S. y TOMKINS, A. Trawling the Web for emerging cyber-communities. *8th. International World Wide Web Conference*, (Toronto, Canada, May 11-14, 1999).
Url: http://www8.org/w8-papers/4a-search-mining/trawling/trawling.html

(Lancaster, 1977) LANCASTER, F. W. *The measurement and evaluation of library services.* Washington D.C.: Information Resources Press, 1977.

(Lancaster, 1991) LANCASTER. F. W. *Bibliometric methods in assessing productivity and impact of reserach.* Bangalore: Sarada Ranganathan Endowment for Library Science, 1991.

(Larson, 1996) LARSON, R. R. Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace. *Annual meeting of the American Society for Information Science*, (Baltimore, October 19-24, 1996), p. 71-78.
Url: http://sherlock.berkeley.edu/asis96/asis96.html

(Lawrence, 1998) LAWRENCE, S. y GILES, C. L. Searching the World Wide Web. *Science*, 1998, Vol. 280, p. 98-100.

(Lawrence, 1999) LAWRENCE, S. y GILES, C. L. Accessibility of information on the Web. *Nature*, 1999, Vol. 400, p. 107-109.

(Leiner, 1997) LEINER, B. M., CERF, V. G., CLARK, D. D., KAHN, R. E., KLEINROCK, L., LYNCH, D. C., POSTEL, J., ROBERTS, L. G. y WOLFF, S. A brief history of the Internet. *On the Internet*, 1997, No. MAyo/Junio.
Url: http://www.isoc.org/internet/history/brief.html

(Leiner, 1997b) LEINER, B. M., CERF, V. G., CLARK, D. D., KAHN, R. E., KLEINROCK, L., LYNCH, D. C., POSTEL, J., ROBERTS, L. G. y WOLFF, S. Una breve historia de Internet (Primera Parte). *Novática*, 1997, No. 130, p. 4-12.

(Leiner, 1998) LEINER, B. M., CERF, V. G., CLARK, D. D., KAHN, R. E., KLEINROCK, L., LYNCH, D. C., POSTEL, J., ROBERTS, L. G. y WOLFF, S. Una breve historia de Internet (Segunda Parte). *Novática*, 1998, No. 131, p. 44-49 .

(Leydesdorff, 1999) LEYDESDORFF, L. y WOUTERS, P. Between Texts and Contexts: Advances in Theories of Citation ? *Scientometrics* , 1999, Vol. 44, No. 2, p. 169-182.
Url: http://www.chem.uva.nl/sts/loet/citation/rejoin.htm

(Licklider, 1962) LICKLIDER, J. C. R. y CLARK, W. On-Line Man Computer Communication. [en línea].[Citado. Disponible en Internet:
http://www-personal.umich.edu/~mattkaz/history/licklider.html

(Lindner, 1994) LINDNER, P. Frequently asked questions about Gopher. [en línea]. 1994 [Citado: Septiembre 1999]. Disponible en Internet:
ftp://rtf.mit.edu/usenet/news.answers/gopher-faq

(Lord, 1984) LORD, S. Le role de l'analyse de ditations dans l`histoire des sciences. *Argus*, 1984, Vol. 13, No. 2, p. 59-65.

(López, 1986) LÓPEZ LÓPEZ, P. *Introducción a la bibliometría.* Valencia: Promolibro, 1986.

(López, 1998) LÓPEZ DE PRADO, R. Museos en Internet: Análisis de recursos documentales. *VI Jornadas Españolas de Documentación, FESABID 98*, (Valencia (España), 29-31 Octubre 1998).
Url: http://www.florida-uni.es/~fesabid98/Comunicaciones/r_lopez/r_lopez.htm

(Malkin, 1993) MALKIN, G. y LAQUEY PARKER, T. Internet Users' Glossary, Network Working Group RFC 1392/FYI 18. [en línea]. January 1993 [Citado: Noviembre 1999]. Disponible en Internet: http://www.cis.ohio-state.edu/htbin/rfc/rfc1392.html

(Martínez, 1989) MARTÍNEZ DE SOUSA, J. *Diccionario de bibliología y ciencias afines.* Salamanca; Madrid: Fundación Germán Sánchez Ruipérez, 1989.

(Mauldin, 1994) MAULDIN, M. y LEAVITT, J. Web agent related reserach at the Center for Machine Translation. *Reunión del ACM Special Interest Gropu on Networked Information Discovery and Retrieval*, (McLean, VA, USA, 4 de Agosto de 1994).
Url: http://www.lazytd.com/lti/pub/signidr94.html

(Mauldin, 1995) MAULDIN, M. Measuring the Web with Lycos. *Proc 3rd. International World-Wide Web Conference*, (Darmstadt, Alemania, 10-14 Abril 1995).
Url: http://www.igd.fhg.de/archive/1995_www95/proceedings/posters/38/

(Mauldin, 1996) MAULDIN, M. Spidering bof report. [en línea]. 1996 [Citado: Enero 2000]. Disponible en Internet:
http://www.w3.org/Search/9605-Indexing-Workshop/ReportOutComes/Spidering.txt

(Mccain, 1990) MCCAIN, K. W. Mapping Authors in intellectual Space: A Technical Overview. *Journal of the American Society for Information Science*, September 1990, Vol. 41, No. 6, p. 433-443.

(Mccain, 1991) MCCAIN, K. W. Mapping Economics through the Journal Literature: An Experiment in Journal Cocitation Analysis. *Journal of the American Society for Information Science*, May 1991, Vol. 42, No. 4, p. 290-296.

(McKiernan, 1996) MCKIERNAN, G. CitedSites(s): Citation Indexing of Web resources. [en línea]. 1996 [Citado: Enero 2000]. Disponible en Internet:
http://www.public.iastate.edu/~CYBERSTACKS/Cited.htm

(McMurdo, 1996) MCMURDO, G. Net by numbers. *Journal of Information Science*, 1996, Vol. 22, No. 5, p. 381-390.

(Medina, 2000) MEDINA, A., MATTA, I. y BYERS, J. On the origin of power laws in internet topologies. *Computer Communication review*, 2000, Vol. 30, No. 2. Url: http://www.cs.bu.edu/techreports/00-004-power-laws-Internet-topology.ps.Z

(Mendelzon, 1995) MENDELZON, A. y WOOD, P. Finding regular simple paths in grapg databases. *SIAM J. Comp.*, 1995, Vol. 24, No. 6, p. 1235-1258.

(Mendelzon, 1997) MENDELZON, G. M. y MILO, T. Querying the World Wide Web. *Journal of Digital Libraries* , 1997, Vol. 1, No. 1, p. 68-88.

(MIDS, 1999) MATRIX INFORMATION AND DIRECTORY SERVICES (MIDS). State of the Internet. [en línea]. 1999 [Citado: Didiembre 1999]. Disponible en Internet: http://www.mids.org/mmq/603/pages.html

(Mukherjea, 1997) MUKHERJEA, S. y HARA, Y. Focus+Context Views of World-Wide Web Nodes. *Proceedings of Hypertext'97*, (Southampton, UK, 1997), p. 187-196.

(Méndez, 1986) MÉNDEZ, A. *Los indicadores bibliométricos. Política científica, 1986. Citado en LÓPEZ LÓPEZ, P. Introducción a la bibliometría*. Valencia: Promolibro, 1986.

(Nalimov, 1969) NALIMOV, V. V. y MULCHENKO, B. M. *Naukometriya.* Moscow: Nauka, 1969.

(Nelson, 1965) NELSON, T. H. A file Structure for the Complex, The Changing and The Indeterminate. *ACM 20th National Conference*, (1965).

(Nicholas, 1978) NICHOLAS, D. y RITCHIE, M. *Literature and bibliometrics.* London: Clive Bingley, 1978.

(Notess, 2000) NOTESS, G. Search Engine Inconsistencies. *Online*, 2000, Vol. 24, No. 2. Url: http://www.onlineinc.com//onlinemag/OL2000/net3.html

(OCLC, 1999) ONLINE COMPUTER LIBRARY CENTER (OCLC). Web characterization project. [en línea]. 1999 [Citado: Diciembre 1999]. Disponible en Internet: http://wcp.oclc.org

(Okubo, 1977) OKUBO, Y. Bibliometric indicators and analysis of reserach systems: methods and examples. *STI Working Papers*, (Paris, 1977).

(Ortega, 1935) ORTEGA Y GASSET, J. Misión del Bibliotecario. *En: Congreso Internacional de Bibliotecas y Bibliografía*, (Madrid, 1935. Actas y Trabajos del II Congreso Internacional de Bibliotecas y Bibliografía. Madrid: Librería de Julián Barbazán, 1935).

(Osareh, 1996) OSAREH, F. Bibliometrics, citation analysis and co-citation analysis: a review of literature I. *Libri*, 1996, Vol. 46, p. 149-158.

(Otlet, 1934) OTLET, P. *Traite de documentation. Le livre sur le livre. Theorie et pratique.* Brussels: Van Keerberghen, 1934.

(Le Pair, 1988) LE PAIR, C. *The citation gap pf applicable science. EN: Handbook of quantitative studies of science and technology.* Amsterdam: North-Holland, 1988.

(Palmer, 2000) PALMER, C. R. y STEFFAN, J. G. Generating Network Topologies That Obey Power Laws. *Proceedings of the Global Internet Symposium, Globecom2000*, (, November 2000, San Francisco).

(Pansiot, 1998) PANSIOT, J. J. y GRAD, D. On routes and multicast trees in the Internet. *ACM Computer Communication Review*, January 1998, Vol. 28, No. 1, p. 41-50.

(Parunak, 1989) PARUNAL, H. V. Hypermedia topologies and user navigation. *Hypertext'89 proceedings*, (Pittsburgh, November 5-8, 1989), p. 43-50.

(Pirolli, 1996) PIROLLI, P., PITKOW, J. y RAO, R. Silk from a Sow's ear: extracting usable structures from the Web. *Conference on Human Factors in Computing Systems, CHI'96*, (Vancouver, April 13-18, 1996).
Url: http://www.acm.org/sigchi/chi96/proceedings/papers/Pirolli_2/pp2.html

(Pitkow, 1998) PITKOW, J. E. Summary of WWW characterizations. *Proceedings for the Seventh International World Wide Web Conference*, (Brisbane, Australia, 14-18 April 1998).
Url: http://www7.conf.au/programme/fullpapers/1877/com1877.htm

(Polanco, 1995) POLANCO, X. Aux Sources de la Scientometrie. *Solaris*, 1995, No. 2.
Url: http://www.info.unicaen.fr/bnum/jelec/Solaris/d02/2polanco1.html

(Potter, 1981) POTTER, W. G. Introduction. *Library Trends*, 1981, Vol. 30, No. 1, p. 5-7.

(Pritchard, 1969) PRITCHARD, A. Statistical bibliography or bibliomrtics? *Journal of Documentation*, 1969, Vol. 25, No. 4, p. 348-349 .

(Quarterman, 1990) QUARTERMAN, J. *The Matrix: Computer Networks and Conferencing Systems WorldWide*. Bedford, MA: Digital Press, 1990.

(Quarterman, 1999) QUARTERMAN, J. S. Internet growth graph. [en línea]. 1999 [Citado: Noviembre 1999]. Disponible en Internet: http://www.mids.org

(Randic, 1975) RANDIC, M. On characterization of molecular branching. *Journal od the American Chemical Society*, 1975, Vol. 97, p. 6609-6615.

(Ravichandra, 1993) RAVICHANDRA RAO, I. K. Guest editorial: librametry to bibliometrics to informetrics ... *Library Science*, 1993, Vol. 30, No. 2, p. i-ii.

(Roberts, 1966) ROBERTS, L. y MERRIL, T. Toward a Cooperative Network of Time-Shared Computers. *Fall AFIPS Conf*, (Oct, 1966).

(Rodríguez, 1997) RODRÍGUEZ I GARÍN, J. M. Valorando el impacto de la información en Internet: Altavista, el "Citation Index" de la Red. *Revista Española de Documentación Científica*, 1997, Vol. 20, No. 2, p. 175-181.
Url: http://escher.upc.es/usr/josep-m/publica/ALTAVIS.HTM

(Rousseau, 1997) ROUSSEAU, R. Sitations: an exploratory study. *Cybermetrics*, 1997, Vol. 1, No. 1.
Url: http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

(Sancho, 1990) SANCHO, R. Indicadores bibliométricos utilizados en la evaluación de la ciencia y la tecnología. Revisión bibliográfica. *Revista Española de Documentación Científica*, 1990, Vol. 13, No. 3-4, p. 842-865.

(Sandison, 1989) SANDISON, A. Documentation note: thinking about citation analysis. *The Journal of Documentation*, 1989, Vol. 45, No. 1, p. 59-64.

(Schrader, 1981) SCHRADER, A. M. Teaching bibliometrics. *Library Trends*, 1981, Vol. 30, No. 1, p. 151-172.

(Sedgewick, 1990) SEDGEWICK, R. *Algorithms in C.* Addison-Wesley, 1990.

(Sengupta, 1992) SENGUPTA. Bibliometrics, Informetrics, Scientometrics and Librametrics: an overview. *Libri*, 1992, Vol. 42, No. 2, p. 75-98.

(Shiode, 2000) SHIODE, N. y BATTY, M. Power law distributions in real and virtual worlds. *INET 2000 Proceedings*, (Yokohama, Japan, 18-21 July 2000).
Url: http://www.isoc.org/inet2000/cdproceedings/2a/2a_2.htm

(Shiri, 1998) SHIRI, A. A. Cybermetrics; a new horizon in information research. *49th FID Conference and Congress*, (New Delhi, India, 11-17 october 1998).

(Small, 1973) SMALL, H. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 1973, Vol. 24, No. 3, p. 265-269.

(Smeaton, 1992) SMEATON, A. F. Information retrieval and hypertext: competing technologies or complementary access methods. *Journal of Information Systems*, 1992, Vol. 2, p. 221-233.

(Smeaton, 1995) SMEATON, A. F. y MORRISEY, P. J. Experiments on the Automatic Construction of Hypertext from Text. *The New Review of Hypermedia and Multimedia: Applications and Research*, 1995, Vol. 1.
Url: http://www.compapp.dcu.ie/~asmeaton/pubs/Hypermedia_Paper.ps

(Smeaton, 1995b) SMEATON, A. F. Building hypertext under the influence of topology metrics. *International Workshop on Hypermedia Design*, (Montpellier, June 1995).
Url: ftp://ftp.compapp.dcu.ie/pub/w-papers/1995/CA0895.ps.Z

(Smith, 1981) SMITH, L. C. Citation analysis. *Library Trends*, 1981, Vol. 30, No. 1, p. 83-106.

(Smith, 1999) SMITH, A. G. A tale of two Web spaces: comparing sites using Web impact factors. *Journal of Documentation*, December 1999, Vol. 55, No. 5, p. 577-592.

(Smith, 1999) SMITH, A. G. ANZAC webometrics: exploring Australasian Web structures. *Proceedings of the Ninth Australasian Information Online & On Disc Conference and Exhibition*, (Sydney. 19–21 January 1999).

Url: http://www.csu.edu.au/special/online99/proceedings99/203b.htm

(Snyder, 1999) SNYDER, H. y ROSENBAUM, H. Can search engines ne used as tools for web-link analysis? A critical view. *Journal of Documentation*, September 1999, Vol. 55, No. 4, p. 375-384.

(Sonnenreich, 1998) SONNENREICH, W. y MACINTA, T. A History of Search Engines. WebDeveloper.com Guide to Search Engines. [en línea]. 1998 [Citado: Enero 2000]. Disponible en Internet: http://www.wiley.com/compbooks/sonnenreich/history.html

(Stevens, 1953) STEVENS, R. E. *Characteristics of subject literature. En: American Colleges & Research Library (ACRL) monographs.* Washington D.C.: Association of College and Reference Libraries a division of the American Library Association, 1953.

(Termens, 1997) TERMENS GRAELLS, M. Les Webs de les biblioteques de Catalunya: estructura interna i enllaços. *6es. Jornades Catalanes de Documentació*, (Barcelona, 23-25 Octubre, 1997), p. 507-517.

(Turnbull, 1996) TURNBULL, D. Bibliometrics and the World-Wide Web. [en línea]. 1996 [Citado: Diciembre 1998]. Disponible en Internet: http://donturn.fis.utoronto.ca/research/bibweb.pdf

(Weider, 1994) WEIDER, C. y DEUTSCH, P. Uniform Resource Names. Internet Draft. IRTF, URI Working Group. [en línea]. 1994 [Citado: Septiembre 1999]. Disponible en Internet: ftp://ftp.isi.edu/internet-drafts/draft-ieft-uri-resource-names-02.txt

(Wheeler, 1999) WHEELER, D. C. y O'KELLY, M. E. Network topology and city accessibility of the Commercial Internet. *Proffesional Geographer*, 1999, Vol. 51, No. 3, p. 327-339.

(White, 1989) WHITE, H. D. y MCCAIN, K. W. Bibliometrics. *Annual review of Information Science and Technology*, 1989, No. 24, p. 119-186.

(White, 1998) WHITE, H. D. y MCCAIN, K. W. Visualizing a discipline: an author co-citation analysis of information science. *Journal of the American Society for Information Science*, 1998, Vol. 49, No. 4, p. 327-355.

(Willet, 1981) WILLETT, P. A fast procedure for the calculation of similarity coefficients in automatic classification. *Information Processing & Management*, 1981, Vol. 17, p.

53-60.

(Woodruff, 1996) WOODRUFF, A. An Investigation of Documents from the World Wide Web. *Fifth International World Wide Web Conference* , (París, May 6-10 1996). Url: http://www5conf.inria.fr/fich_html/papers/P7/Overview.html

(Zhang, 1999) ZHANG, J. y KORFHAGE, R. R. DARE: distance and angle retrieval environment: a tale of the two measures. *Journal of the American Society for Information Science*, 1999, Vol. 50, No. 9, p. 779-787.

(Zhang, 1999b) ZHANG, J. y KORFHAGE, R. R. A distance and angle similarity measure method. *Journal of the American Society for Information Science*, 1999, Vol. 50, No. 9, p. 772-778 .

(Zipf, 1949) ZIPF, G. K. *Human Behavior and the Principle of Least Effort*. Cambridge, Massachusetts: Addison-Wesley, 1949.